Offer #2024-07755

PhD Position F/M Incremental Deep Learning for Embedded Systems

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Other valued qualifications : Computer Science, Neuroscience, Math, Statistic

Fonction : PhD Position

About the research centre or Inria department

The Inria Rennes - Bretagne Atlantique Centre is one of Inria's eight centres and has more than thirty research teams. The Inria Center is a major and recognized player in the field of digital sciences. It is at the heart of a rich R&D and innovation ecosystem: highly innovative PMEs, large industrial groups, competitiveness clusters, research and higher education players, laboratories of excellence, technological research institute, etc.

Context

Context

This PhD will occur in the context of the project Adapting (https://www.pepr-ia.fr/en/projet/adapting-2/) from
the PEPR AI (https://www.pepr-ia.fr/en/pepr/). This project focuses on designing adaptive embedded hardware architectures for AI. In this context, our team wants to design new incremental machine learning
algorithms that could serve as use cases in the Adapting project for other researchers who will focus on the
hardware architecture design.
Continual learning , also known as lifelong learning or incremental learning, is a machine learning paradigm that focuses on the ability of a model to learn and adapt continuously over time as it encounters
new data, tasks, or concepts, without forgetting or catastrophically overwriting what it has previously learned.
In continual/incremental learning, the learned model should retain knowledge about previous tasks or data
while incorporating new information. In this PhD, we will focus on designing new resource-efficient incremental learning algorithms that can run on embedded systems with their associated ressource and privacy constraints. These contraints involve limited computational power, memory, and energy efficiency.
They also involve real-time processing with low latency and often deterministic behavior. Updating embedded models is complex due to hardware limitations and the need for efficient updates while handling
data locally to enhance privacy and security.
This PhD will focus on foundation models such as well-known LLM -Large Language Models- (e.g. GPT-3.5,
Mixtral, Llama 3,...) or multimodal ones (involving for example ViT -Vision Transformer- models such as GPT-4o, Sora, Dall-E 3) and their ability to evolve continuously in an embedded environnement.

Application process
The position is funded for 3 years (this is the standard duration of a PhD in France). The net salary is around
2000 euros. The PhD student will be based in Rennes (https://en.wikipedia.org/wiki/Rennes) and will make
a few stays in Grenoble during the 3-years contract.
Applications will be processed on a first-come, first-served basis until June 15, 2024.

Application Material and Procedure
Here is the supervision team:

- Denis Coquenet, Associate professor, Université de Rennes (denis.coquenet@irisa.fr).
- Elisa Fromont, Professor, Université de Rennes (elisa.fromont@irisa.fr).
- Martial Mermillod, Professor (in Cognitive Sciences) UGA. MIA chair on "Core AI-Artificial Neural Networks" (martial.mermillod@univ-grenoble-alpes.fr).
- Marina Reyboz, PhD, CEA (marina.reyboz@cea.fr).

Applicants should send these documents to the *entire* supervision team :

- An academic CV.
- An official transcript listing classes taken and grades received.
- A short statement (maximum of 2 pages) that explains what your research interests are and why you would like to work with us.
- A copy of your most recent English report (e.g. your master report).
- Any published and relevant research papers (in English, preferably PDF format). If you have several papers, please indicate which ones you consider the three most important ones.
- A list of references (e.g., supervisors)

Assignment

Related work

Incremental learning consists in a multi-stage training in which the context (data domain, classes, or task)
evolves between each training stage. This paradigm faces the stability-plasticity dilemma. It means that the
model must continuously adapt to new contexts while avoiding catastrophic forgetting (performance on past
contexts must not deteriorate). The naive approach is the standard fine-tuning strategy in which the parameters of the model (or part of it) are updated from one stage to another by training on the new context
only. It is known to be efficient for adaptation but it is prone to catastrophic forgetting. Several works focused
on replay strategies to tackle this challenge. It consists in managing a buffer of examples from past contexts, which are preserved through the different training stages. Alternatives have been
proposed to avoid storing old data, for instance by learning to preserve the latent representation on the new
data through the training stage . However, current incremental foundation models do not include important
desiderata for embedded models (e.g. lack of privacy, restricted hardware resources and low energy consumption, robustness, no need for an oracle or neurogenesis). Our current project will explore the models
[1, 2, 3,4,5,6,7] and methods having those important requirements for edge AI. Another line of research is the parameter-oriented approaches: the goal is to associate part of the model's parameters with a specific context in order to maximize the performance on both past and current contexts. This is mainly done by mapping parameters
to contexts , parameter pruning or parameter addition . Among the parameter-oriented techniques, a recent trend focused on prompt tuning, used for transformer architectures . It consists in learning input tokens which are prepended to the input to condition the behavior of the model whose weights are frozen.
While parameter pruning and mapping lead to saturation after many training stages, the parameter addiction
approach lead to a growing number of parameters.
A complementary approach is the knowledge-editing strategy which aims at correcting specific knowledge by locating and updating the responsible neurons.

Bibliography

1. Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, Tinne Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks", IEEE Trans. Pattern Anal. Mach. Intell (TPAMI), 44:7, 2022.
2. Li et al., "Learning without Forgetting", European Conference on Computer Vision (ECCV) 2016 https://arxiv.org/pdf/1606.09282.pdf
3. Chaudhry et al., "Continual learning with tiny episodic memories", preprint 2019 https://arxiv.org/pdf/1902.10486
4. Serra et al., “Overcoming catastrophic forgetting with hard attention to the task”, International Conference on Machine Learning (ICML), 2018 https://arxiv.org/pdf/1801.01423.pdf
5. Mallya et al., "PackNet: Adding Multiple Tasks to a Single Network by Iterative Pruning", Conference on Computer Vision and Pattern Recognition (CPPR), 2018 https://openaccess.thecvf.com/content_cvpr_2018/papers/Mallya_PackNet_Adding_Multiple_CVPR_2018_paper
6. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models", International Conference on Learning Representations (ICLR), 2022https://arxiv.org/pdf/2106.09685.pdf
7. Lester et al., "The Power of Scale for Parameter-Efficient Prompt Tuning", Conference on Empirical Methods in Natural language Processing (EMNLP), 2021https://arxiv.org/pdf/2104.08691.pdf
8. Meng et al., "Mass-editing Memory in a Transformer", International Conference on Learning representations (ICLR), 2023

Main activities

Objectives

The objective of this PhD is to investigate incremental learning of foundation models on embedded systems.
The graduate student will initially address the question: "Are foundation models prone to catastrophic forgetting with standard fine-tuning techniques?" by conducting comprehensive studies and experiments.

Subsequently, she/he will explore how state-of-the-art (SOTA) methods can be applied to these models, assessing the extent to which these existing techniques provide effective solutions to the problem of incremental learning. This phase will involve adapting and optimizing SOTA methods to suit the constraints
and requirements of embedded systems (efficiency, privacy, low energy consumption, robustness, etc.). Finally, the PhD student will develop and propose new methodologies specifically designed to enable foundation models to learn incrementally when deployed on embedded systems, ensuring that these models
maintain performance and adaptability over time.

What we are looking for
The ideal candidate will possess the following skills and attitudes:

- A master degree in computer science, AI, or a closely related discipline.
- A strong background in computer science and mathematics.
- A scientific attitude and the ability to reason through problems.
- Excellent programming skills.
- The ability to communicate written and orally in English in a clear and precise manner.
- A pro-active and independent attitude as well as the ability to function well in a team environment.
- A good motivation for pursuing a Ph.D. and working in Rennes or Grenoble.

Skills

(see ideal candidate before)

Benefits package

- Subsidized meals
- Partial reimbursement of public transport cost

Remuneration

monthly gross salary amounting to 2100 euros

General Information

- **Theme/Domain :** Data and Knowledge Representation and Processing
  Statistics (Big data) (BAP E)
- **Town/city :** Rennes
- **Inria Center :** Centre Inria de l'Université de Rennes
- **Starting date :** 2024-10-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2024-07-26

Contacts

- **Inria Team :** LACODAM
- **PhD Supervisor :**
  Fromont Elisa / elisa.fromont@irisa.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.

The keys to success

What we are looking for
The ideal candidate will possess the following skills and attitudes:

- A master degree in computer science, AI, or a closely related discipline.
- A strong background in computer science and mathematics.
- A scientific attitude and the ability to reason through problems.
- Excellent programming skills.
- The ability to communicate written and orally in English in a clear and precise manner.
- A pro-active and independent attitude as well as the ability to function well in a team environment.
- A good motivation for pursuing a Ph.D. and working in Rennes or Grenoble.

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

Please submit online : your resume, cover letter and letters of recommendation eventually

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.