Postdoctoral Research

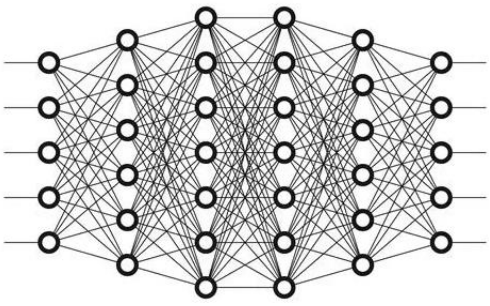
Reconfigurable AI Systems : Dynamically Adaptable Neural Networks

Bogdan Robu (bogdan.robu@gipsa-lab.fr) **Eric Rutten (eric.rutten@inria.fr)**
Matteo Tacchi (matteo.tacchi@grenoble-inp.fr)

Context : https://project.inria.fr/radyal/

Research on machine learning and (Deep) Neural Networks (DNN) has made considerable progress in the past decades. State-of-the-art DNN models usually require large amounts of data to be trained and contain a tremendous number of parameters leading to overall high resource requirements, in terms of computation and memory and thus energy [6]. The desire to **reach out Digital Sobriety**, gave rise to approaches which try to reduce these requirements and, during or after training, remove parts of the DNN model (pruning) [2], store it with lower precision (quantization) [4], train surrogate models (knowledge distillation) or search the best configuration by trying different parameters (Neural Architecture Search).

Concerning the hardware, many optimizations have been proposed to accelerate the inference of DNNs on different architectures mainly FPGA [1][3] or Cloud execution [5]. But these accelerators are usually specific to a given hardware or software configuration (– a DNN in our case) and are optimized to satisfy certain static performance criteria. However, for many applications, the **performance requirements of a DNN model deployed on a hardware platform are not static but evolving dynamically** because of unwanted behavior due to faults [9] or as its operating conditions and environment change [8]. In this case the DNN needs to be able to somehow adapt itself [7] and become more compact [10] without much loss in performance..



Thus, in this project we propose an original interdisciplinary approach that allows **DNN models to be dynamically (re)configurable** at run-time on a given reconfigurable hardware accelerator architecture, depending on the external environment, following an approach based on *feedback loops and control theory*.

Postdoctoral Research

This research project is a collaboration between University Grenoble Alpes through the GIPSA-lab research laboratory for their expertise in control engineering, INSA Lyon through the LIRIS laboratory (for their expertise on DNN conception) and Inria Research center in Rennes (for their expertise on FPGA implementation).
This research is part of ***RADYAL*** *(Resource-Aware DYnamically Adaptable machine Learning)* project financed through the ANR call TSIA – Specific Research in Artificial Intelligence:
https://project.inria.fr/radyal/

Expected work

This work is at the intersection of the work of our colleagues in Lyon and Rennes !
Some ideas about the path to follow:

- Use the first results from Rennes and Lyon to develop two different control algorithms: one for the varying DNN model with a fixed Hardware (HW) architecture and another for different HW configurations running a fixed DNN model. On one side, the first algorithm, will use the space of possible run-time configurations as well as the control parameters of the DNN model, obtained to introduce a first control algorithm that will be capable to drive a DNN model on a fixed HW configuration in order to achieve a desired objective (increase performance, reduce energy consumption, etc.). On the other side, the second algorithm, for a given static DNN model, will use the possible HW combinations to achieve the desired objectives. These algorithms will have the capacity to auto-tune themselves at runtime.
- Combine the two previous algorithms to achieve a global controller capable of dealing with the DNN and the HW while coordinating the two loops. Different techniques will be proposed based on the application use case: different priority orders of criteria as well as different priority orders between HW and Software loops.
- Implement, test and evaluate the proposed solution in cooperation with our partners.
  Help for the implementation on the actual systems will be provided by the engineers in Rennes and Lyon.

Required skills

We seek a talented and motivated postdoc candidate passionate about research, practical implementation issues and with good English and programming skills in Python.
The candidate should have either a solid background in control theory and knowledge in neural networks programming OR a solid background in the programming of neural networks and good knowledge in applied mathematics !

Knowledge on FPGA implementation is an added value but is not mandatory as the actual implementation will be done by the engineers of our research team.
NO teaching is required ! If the Post-Doctoral researcher wants to teach (in French or English) the taught hours are paid separately by the university!

Location: The project will be carried out in the GIPSA-lab research laboratory of Université Grenoble Alpes. Regular visits can be carried out in Lyon and Rennes for project meetings and interactions with the research teams.

Université Grenoble Alpes is a multidisciplinary institution located in the Heart of the French Alps, renowned for its scientific and technological research activities. Rooted in its territory,

GIPSA-lab
Campus universitaire
961 rue de la Houille Blanche - BP46
F-38402 GRENOBLE Cedex

www.gipsa-lab.fr

bogdan.robu@gipsa-lab.fr

multidisciplinary and open to the world, counts 59,000 students, including 10,500 international students, 3,000 PhD students, and 7,800 staff.

What we offer:

Package for hiring 2-3 master students during the 24 months.
Possibility to teach at the University, in which case the hours taught are paid on top of the salary !
Our working time agreement to preserve your work-life balance, with several days off (paid leave, etc), fixed and flexible working hours as well as the possibility of working from home.
Our work environment: co-working space, gym, sport clubs on the campus, green working sites

Deadline: 15th of December, 2024
Duration: 24 months
Net monthly salary (net income after taxes) : 2500 € / month

Application information

Please send your detailed CV, including a list of publications, motivation letter, and contact details for one or two references to: Bogdan Robu, Eric Rutten and Matteo Tacchi.

Keywords: Online training, Feedback control, Neural networks, Reconfiguration, Digital Sobriety, FPGA

References

1. WANG, C.; LUO, Z. A Review of the Optimal Design of Neural Networks Based on FPGA. Appl. Sci. 2022, 12, 10771. https://doi.org/10.3390/app122110771
2. Niu, Wei, et al. "Patdnn: Achieving real-time dnn execution on mobile devices with pattern-based weight pruning." Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems. 2020.
3. Marcello Traiola, Fernando Fernandes dos Santos, Paolo Rech, Carlo Cazzaniga, Olivier Sentieys, et al.. Impact of High-Level-Synthesis on Reliability of Artificial Neural Network Hardware Accelerators. IEEE Transactions on Nuclear Science, 2024, pp.1-9.
4. Fox, Sean, et al. "Training deep neural networks in low-precision with high accuracy using FPGAs." 2019 International Conference on Field-Programmable Technology (ICFPT). IEEE, 2019.
5. CHEN, Yao, HE, Jiong, ZHANG, Xiaofan, et al. Cloud-DNN: An open framework for mapping DNN models to cloud FPGAs. In : Proceedings of the 2019 ACM/SIGDA international symposium on field-programmable gate arrays. 2019. p. 73-82.
6. A Berthelot, E Caron, M Jay, L Lefèvre. Estimating the environmental impact of Generative-AI services using an LCA-based methodology - Procedia CIRP, 2024
7. Zilong Zhao, Sophie Cerf, Bogdan Robu, Nicolas Marchand. Event-Based Control for Online Training of Neural Networks. IEEE Control Systems Letters, 2020, 4 (3), pp.773-778.
8. Zilong Zhao, Sophie Cerf, Bogdan Robu, Nicolas Marchand. Feedback Control for Online Training of Neural Networks. CCTA 2019 - 3rd IEEE Conference on Control Technology and Applications, Aug 2019, Hong Kong, China.
9. Marcello Traiola, Salvatore Pappalardo, Ali Piri, Annachiara Ruospo, Bastien Deveautour, et al.. Approximate Fault-Tolerant Neural Network Systems. ETS 2024 - 29th IEEE European Test Symposium, May 2024, La Haye, Netherlands. pp.1-10.
10. Anthony Berthelier, Yongzhe Yan, Thierry Chateau, Christophe Blanc, Stefan Duffner, et al.. Learning Sparse Filters In Deep Convolutional Neural Networks With A l 1 /l 2 Pseudo-Norm. CADL 2020 : Workshop on Computational Aspects of Deep Learning - ICPR 2020, Jan 2021, Milan, Italy.

GIPSA-lab
Campus universitaire
961 rue de la Houille Blanche - BP46
F-38402 GRENOBLE Cedex

www.gipsa-lab.fr

bogdan.robu@gipsa-lab.fr

https://project.inria.fr/radyal/