Call for applications – PhD candidate

Lyon Institute of Nanotechnology
Ecole Centrale de Lyon, 36 av. Guy de Collongue,
F-69134 Ecully, FRANCE
http://inl.cnrs.fr

Extreme compression schemes for edge AI

Deep Neural Networks (DNNs) [1] are currently one of the most intensively and widely used predictive models in the field of machine learning. DNNs have proven to give very good results for many complex tasks and applications, such as object recognition in images/videos, natural language processing, satellite image recognition, robotics, aerospace, smart healthcare, and autonomous driving. Nowadays, there is intense activity in designing custom Artificial Intelligence (AI) hardware accelerators to support the energy-hungry data movement, speed of computation, and memory resources that DNNs require to realize their full potential [2]. Furthermore, there is an incentive to migrate AI from the cloud into the edge devices, i.e., Internet-of-Things (IoTs) devices, in order to address data confidentiality issues and bandwidth limitations, given the ever-increasing internet-connected IoTs, and also to alleviate the communication latency, especially for real-time safety-critical decisions, e.g., in autonomous driving.

As pointed out in [3], one of the major issues is the memory required to store the DNN parameters (i.e., weights and bias). Indeed, higher the memory footprint higher the energy and latency required to transfer the data from/to the storage to the computational unit. Moreover, for some edge devices the on-board memory is simply too small to contain all DNN parameters. To alleviate this problem, several techniques explored the precision reduction of the data representation of DNN parameters. **Quantization** is one of the well-known and most popular techniques, it consists of using integer data type (16, 8, 4 or even 1 bit) during the inference instead of the 32-bit floating-point data used for training. Another technique is the **Weight-Sharing (WS)** [4]. WS aims at clustering together DNN parameters having similar values. WS can be applied before/after quantization and [4] shown that it is possible to achieve up 5x memory foot print reduction. A different approach has been presented in [5], in which a regression technique is used to generate (infer) the set of DNN parameters. Only a subset of parameters (i.e., 10%) is required to be stored and the regression leverages on this subset to generate the full set of DNN parameters with a certain degree of precision.

This project aims at investigating the concept of **extreme** compression of DNN parameters by leveraging on multiple techniques. Instead of regression, we intend to study the concepts of **lossless data compression, autencoders** [6] and **generative AI** [7] to be able to run DNN is a very constrained environment (i.e., with a RAM of capacity varying from KB to few MB) such as ultra-low power microcontrollers.

In the framework of the French research project AdaptING, the Electronic group at INL will work in collaboration with LIRMM. In this context we are currently looking for a (m/f) **PhD student** for a **3-years**.

Job description

The Ph.D. thesis is structured in the following 4 main tasks

1. **Compression techniques State of the Art and DNN profile (M1 to M9)**
    - Analysis of the compression techniques (e.g., Huffman code, autencoders, generative AI, …)
    - Identification of DNN case study and profile of the related parameters. The goal is to analyze the data distribution to understand what are the most promising compression technique to be used.
2. **Compression Framework (M9 to M18)**
    - A framework implementing the compression technique has to be developed. The input of the framework will be the set of data to be compressed, the output will be the compressed data.
    - Metrics used to evaluate the compression framework are the compression ratio (output size/input size) and the distance of the decompressed data w.r.t. the input data.
3. **DNN compression architecture (M18-M30)**
    - The compression scheme has to be implemented in order to be used during the inference. The scheme can be implemented as software, or hardware. This task has to investigate implementation solutions in order to reduce the overhead as much as possible.

4. **Evaluation and Dissemination (M12-M36)**

- Accuracy, Memory Footprint, Power Consumption, Performances;
- Evaluation can be done through simulation, microcontroller board measurements, prototypes on FPGA;
- Scientific papers and thesis manuscript preparation.

Profile

You have or are about to obtain an MSc in Computer Engineer / Computer Science with strong experience in at least one of the following areas: computer architectures, digital circuit design, optimization algorithms. Good programming skills (python, C and C++) are required. Previous experience in Neural Networks is a plus (e.g., knowledge of major NN frameworks such as Pytorch and Tensorflow). Excellent written and verbal communication skills in English. Fluency in French is also a plus but is not mandatory.

About INL

INL is a 250-strong research institute based in Lyon, France, carrying out fundamental and applied research in electronics, semiconductor materials, photonics and biotechnologies. The Electronic group is a leader in the area of advanced nanoelectronic design, with research projects and collaborations at both national and European level.

Dates:

Ph.D. will start in 2024.

Environment:

The Ph.D. candidate will be supervised by the INL team in Lyon (Ecole Centrale Campus). The Ph.D. salary will follow standard French rates.

References:

[1] Y. LeCun, et al., "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
[2] B. Moons, et al, "14.5 Envision: A 0.26-to-10TOPS/W subword-parallel dynamic-voltage-accuracy-frequency-scalable Convolutional Neural Network processor in 28nm FDSOI," in IEEE ISSCC, 2017.
[3] Dupuis, E., Filip, S., Sentieys, O., Novo, D., O'Connor, I., Bosio, A. (2022). Approximations in Deep Learning. In: Bosio, A., Ménard, D., Sentieys, O. (eds) Approximate Computing Techniques. Springer, Cham. https://doi.org/10.1007/978-3-030-94705-7_15
[4] E. Dupuis, D. Novo, I. O'Connor, A. Bosio, "A Heuristic Exploration of Retraining-free Weight-Sharing for CNN Compression", in *2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 134-139.
[5] Juza, T., Sekanina, L. (2023). GPAM: Genetic Programming with Associative Memory. In: Pappa, G., Giacobini, M., Vasicek, Z. (eds) Genetic Programming. EuroGP 2023. Lecture Notes in Computer Science, vol 13986. Springer, Cham. https://doi.org/10.1007/978-3-031-29573-7_5
[6] U. Michelucci, "An Introduction to Autoencoders." arXiv, 2022. doi: 10.48550/ARXIV.2201.03898.
[7] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, "Generative AI," Business & Information Systems Engineering, vol. 66, no. 1. Springer Science and Business Media LLC, pp. 111–126, Sep. 12, 2023. doi: 10.1007/s12599-023-00834-7.

Send CV and statement of purpose (in English or French) to

Alberto Bosio / INL - Lyon Institute of Nanotechnology - Ecole Centrale Lyon – email: alberto.bosio@ec-lyon.fr