Ph.D. Position:
Ultra-Low Energy Inference Using Redundant Neural Networks

Summary

Research areas:	Machine learning, VLSI system design
Supervisors:	Dr. Vincent Gripon (IMT Atlantique; vincent.gripon@imt-atlantique.fr)
	Dr. François Leduc-Primeau (ETS; francois.leduc-primeau@lacime.etsmtl.ca)
Director:	Prof. François Gagnon (ETS)
Start date:	September 2018 or later.
Institution:	Dual-diploma Ph.D. at École de Technologie Supérieure (Montréal, Canada) and IMT Atlantique (Brest, France).

Context of the Research

Machine learning algorithms based on deep neural networks (DNNs) are achieving outstanding results on many signal processing tasks such as speech recognition and computer vision, among many others. However, even though the prediction (inference) phase of these algorithms is much simpler than the learning phase that must be completed beforehand, it still consumes too much energy to be executed directly on embedded systems. As a result, current systems are forced to transmit the input data to servers in the cloud that perform the inference and return the result, thus greatly increasing the processing latency and the amount of data transmitted by the device. This double penalty prevents the use of DNNs for real-time processing in low-power devices, and also creates privacy issues.

Some recent research has been looking at ways of reducing the computational complexity of inference on DNNs by reducing the size of the network [1] or replacing floating-point operations by binary ones [2, 3]. These contributions are very useful, but in order to obtain the best energy efficiency, it is necessary to combine these algorithm-level improvements with the latest advances in energy-efficient integrated circuits (ICs). Unfortunately, ICs that achieve extreme energy efficiency are usually not very reliable. For instance, CMOS circuits operated in the near-threshold regime can reduce energy consumption by about 10× but are affected by timing variations and memory faults [4]. Other novel technologies such as logic-in-memory devices also come with a similar reliability-energy tradeoff [5].

Objective

The objective of this Ph.D. is to propose approaches to dramatically reduce the energy required by the inference phase of DNN algorithms, by working both on the architecture of the deep neural network and on the hardware implementation of the inference phase, targeting energy-efficient but unreliable integrated circuit technologies. A first aspect of this research will be to study the circuit architectures used to implement state-of-the-art inference processors and characterize how faults occurring in the circuit affects the inference computations. Based on this characterization, the second aspect of the research will be to propose neural network architectures that are better adapted to energy-efficient but faulty integrated circuit technologies.

Supervision and Funding

The research training of the Ph.D. student will be split between *IMT Atlantique* in France and *École de Technologie Supérieure* (ETS) in Canada, and will lead to a double diploma awarded jointly by both institutions. The amount of time spent at each institution can be adjusted based on the preference of the candidate.

At IMT Atlantique, the student will be supervised by Dr. Vincent Gripon. Dr. Gripon's research focuses on machine learning theory and applications. He received his M. Sc. from ENS-Cachan in 2008 and Ph.D. from Télécom Bretagne in 2011. Since then, he coauthored about 70 papers in the fields of artificial neural networks and signal processing.

At ETS, the student will be supervised by Dr. François Leduc-Primeau. Dr. Leduc-Primeau's research focuses on the efficient hardware implementation of communication and signal processing algorithms. Dr. Leduc-Primeau received his Ph.D. from McGill University in 2016. He was awarded several scholarships, including the Postdoctoral Fellowship from Canada's Natural Science and Engineering Research Council (NSERC).

General supervision will be provided by Prof. François Gagnon. Prof. Gagnon has been a professor in the Department of Electrical Engineering at ETS since 1991, and served as the department's director from 1999 to 2001. He is the holder of two research chairs: the Richard J. Marceau Industrial Research Chair for Wireless Internet in developing countries and the NSERC-Ultra Electronics Chair in Wireless Emergency and Tactical Communication. Professor Gagnon serves on the boards of funding agencies and companies. He holds 7 patents and has coauthored more than 250 papers in the fields of wireless communications, signal processing, and microelectronics.

The position is fully funded and travel fees between France and Canada are also covered.

Candidate Profile

- Master's degree in Computer/Electrical Engineering, Computer Science, or related discipline.
- At least one of: 1) prior research experience in machine learning, or 2) prior research experience in VLSI system design.
- Demonstrated programming experience.

How to Apply

Send an e-mail to vincent.gripon@imt-atlantique.fr and francois.leduc-primeau@lacime.etsmtl.ca with the subject line "Ultra-low energy NN PhD position", including a full CV, university transcripts, recommendation letters or contacts from former teachers/advisors, and a statement (max. 1 page) describing how your experience prepares you for this project.

Deadline for applying is April 15th at midnight, time of Montréal.

References

[1] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1mb model size," *CoRR*, vol. abs/1602.07360, 2016. [Online]. Available: http://arxiv.org/abs/1602.07360

[2] M. Courbariaux and Y. Bengio, "Binarynet: Training deep neural networks with weights and activations constrained to +1 or -1," *CoRR*, vol. abs/1602.02830, 2016. [Online]. Available: http://arxiv.org/abs/1602.02830

[3] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: ImageNet classification using binary convolutional neural networks," *CoRR*, vol. abs/1603.05279, 2016. [Online]. Available: http://arxiv.org/abs/1603.05279

[4] M. Alioto, "Ultra-low power VLSI circuit design demystified and explained: A tutorial," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 59, no. 1, pp. 3–29, 2012.

[5] S. Ganapathy, A. Teman, R. Giterman, A. Burg, and G. Karakonstantis, "Approximate computing with unreliable dynamic memories," in *2015 IEEE 13th International New Circuits and Systems Conference (NEWCAS)*, June 2015, pp. 1–4.