UMR IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTEMES ALEATOIRES
Campus universitaire de Beaulieu ● 263 Avenue du Général Leclerc - 35042 Rennes, France
Tél. : +33 2 99 84 71 00 ● Télécopie : +33 2 99 84 75 22 ● Internet : www.irisa.fr
Campus universitaire ENSSAT Lannion ● 6 rue de kerampont – 22305 Lannion, France
Tél. : +33 2 96 46 90 00 ● Télécopie : +33 2 96 46 90 75 ● Internet : www.enssat.fr

Master Thesis proposal

Title: CNN inference approximation: improving performance of deep-learning accelerators

Keywords: deep learning, convolutional neural networks, floating-point, fixed-point, high-level synthesis, hardware acceleration, accuracy analysis, embedded vision systems

Supervisor: Olivier Sentieys (olivier.sentieys@inria.fr), Silviu-Ioan Filip (silviu.filip@inria.fr)
Laboratory: IRISA/Inria – Cairn team
Place: Rennes (Campus de Beaulieu)

Deep Learning [LeCun15], and in particular Convolutional Neural Networks (CNNs), are currently one of the most intensively and widely used predictive models in the field of machine learning. CNNs have shown to give very good results for many complex tasks such as object recognition in images/videos, drug discovery, natural language processing, autonomous driving up to playing complex games [Deng13, Kri12, Chen15, Silv16].

In spite of these benefits, the **computational workload involved in CNNs** is often out of reach for low-power embedded devices, and/or is still very costly when running on datacenter hardware platforms. For example, the amazing performance of AlphaGo [Silv16] required 4 to 6 weeks of training executed on 2000 CPUs and 250 GPUs for a total of about 600kW of power consumption (while the human brain of a go player requires about 20W). Thus, a lot of research effort from both industrials and academics has been concentrated in defining/designing custom hardware platforms supporting this type of algorithms, with the goal of improving performance and/or energy efficiency [Chao17, Hsin17, Zhi17].

In the recent years, **Approximate Computing (AxC)** has become a major field of research to improve both speed and energy consumption in embedded and high-performance systems [Mit16]. By relaxing the need for fully precise or completely deterministic operations, approximate computing substantially **improves energy efficiency**. Various techniques for approximate computing augment the design space by providing another set of design knobs for performance-accuracy tradeoffs. For instance, the gain in energy between a low-precision 8-bit operation suitable for vision and a 64-bit double-precision floating-point operation necessary for high-precision scientific computation can reach up to 50x by considering storage, transport and computation. The gain in energy efficiency (the number of computations per Joule) is even higher. The challenge is then to find adequate (no more and no less) number representations and word-lengths of data/computations that are compatible with application constraints.

CNNs show **inherent resilience to insignificant errors** due to their iterative nature and learning process. Therefore, an intrinsic tolerance to inexact computation is evidenced, and using the approximate computing paradigm to improve power and delay characteristics is therefore relevant [Sung15]. Indeed, CNNs lend well with AxC techniques, especially with fixed-point arithmetic or low-precision floating-point implementations. They are therefore ideally suited for hardware acceleration using FPGA and/or ASIC implementation as acknowledged by the large body of work on this topic.

The goal of this thesis is to explore how approximation techniques can improve the performance of CNNs in deep-learning applications. In particular, we will study how custom floating-point and fixed-point arithmetic, adequate number representations, and even algorithmic-level transformations, can improve performance and energy efficiency of CNN computation while keeping classification and learning phases at a very high success rate. The aim is to go further than current state-of-art research

studies in which only the inputs and outputs of the neural network layers are quantized to low precision.

We will investigate the tradeoffs and benefits that different numeric formats and arithmetic operators bring in terms of performance and energy efficiency when doing Deep Neural Network inference. The goal is for the candidate to contribute to the development of a framework for automatic precision exploration and weight value quantization for CNN inference suited for hardware acceleration using FPGAs.

One particular direction of study we might pursue is that of numerical format optimization when doing inference in network architectures based on structured block-circulant weight matrices [Ding17]. Such architectures offer the promise of small storage requirements (i.e., linear in the number of neurons of each layer vs quadratic in classic architectures) and improved performance in training and inference (i.e., quasi-linear vs quadratic cost in the classic setting), critical elements needed for a wide adoption of deep learning methods on low-power embedded devices.

The Thesis is funded by the ANR *AdequateDL* (Approximating Deep Learning Accelerators) project starting early 2019. The ambition of *AdequateDL* is to explore this new way to get order-of-magnitude improvements in performance and energy efficiency and therefore to influence the design of future computing systems dedicated to deep learning applications in both embedded and cloud markets.

Required or appreciated skills

- mathematical maturity (familiarity with calculus and linear algebra)
- programming experience in C/C++
- familiarity with computer architecture, computer arithmetic, hardware and embedded systems design
- notions of numerical computing (e.g. finite precision computing effects) and/or machine learning would be a plus

References

[Aim17] A. Aimar, H. Mostafa, E. Calabrese, A. Rios-Navarro, R. Tapiador- Morales, I.A. Lungu, M.B. Milde, F. Corradi, A. Linares-Barranco, S.C. Liu, and T. Delbruck, “NullHop: A Flexible Convolutional Neural Network Accelerator Based on Sparse Representations of Feature Maps”, arXiv preprint arXiv:1706.01406, 2017.
[Bar17] B. Barrois and O. Sentieys., “Customizing Fixed-Point and Floating-Point Arithmetic - A Case Study in K-Means Clustering,” In IEEE International Workshop on Signal Processing Systems (SiPS), page 6, October 2017.
[Chao17] W. Chao, L. Gong, Q. Yu, X. Li, Y. Xie, and X. Zhou, “DLAU: A scalable deep learning accelerator unit on FPGA”, IEEE Trans. on Computer-Aided Design of Int. Circ. and Syst., vol. 36, no. 3, 2017, pp. 513-517.
[Chen15] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, “Deepdriving: Learning affordance for direct perception in autonomous driving,” in IEEE International Conference on Computer Vision, 2015, pp. 2722–2730.
[Chen17] Y.H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy- efficient reconfigurable accelerator for deep convolutional neural networks", IEEE Journal of Solid-State Circuits, vol. 52, no. 1, pp. 127- 138, 2017.
[Deng13] L. Deng et al., “Recent advances in deep learning for speech research at microsoft”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 8604–8608.

[Ding17] CirCNN: accelerating and compressing deep neural networks using block-circulant weight matrices. In Proceedings of the IEEE/ACM International Symposium on Microarchitecture (MICRO), 395-408, 2017.
[Guo16] K. Guo, S. Lingzhi, J. Qiu, S. Yao, S. Han, Y. Wang, and H. Yang, "Angel-eye: A complete design flow for mapping cnn onto customized hardware", IEEE Computer Society Annual Symposium on VLSI (ISVLSI), pp. 24-29, July 2016.
[Gys16] P. Gysel, “Ristretto: Hardware-Oriented Approximation of Convolutional Neural Networks”, arXiv preprint arXiv:1505.06402, 2016.
[Han16] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, “EIE: efficient inference engine on compressed deep neural network”, ACM/IEEE 43rd International Symposium on Computer Architecture, pp. 243-254, June 2016.
[Hsin17] C. Yu-Hsin, T. Krishna, J. S. Emer, and V. Sze, “Eyeriss: An energy-efficient reconfigurable accelerator for deep conv. neural networks”, IEEE Journal of Solid-State Circuits, vol. 52, no. 1, 2017, pp. 127- 138.
[Kri12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, Advances in Neural Information Processing Systems, 2012, pp. 1097-1105.
[LeCun15] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning”, Nature, vol. 521, no. 7553, 2015, pp. 436–444.
[Li17] H. Li, S. De, Z. Xu, C. Studer, H. Samet, T. Goldstein, Training Quantized Nets: A Deeper Understanding, Advances in Neural Information Processing Systems (NIPS), pp. 5811--5821, 2017.
[Mit16] S. Mittal, “A Survey of Techniques for Approximate Computing”, ACM Computing Surveys (CSUR), Vol. 48, no. 4, 2016, pp. 1-33.
[Qiu16] J. Qiu, J. Wang, S. Yao, K. Guo, B. Li, E. Zhou, J. Yu, T. Tang, N. Xu, S. Song, Y. Wang, "Going deeper with embedded fpga platform for convolutional neural network", ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA), pp. 26- 35, February 2016.
[Rub13] C. Rubio-Gonzalez et al., “Precimonious: Tuning assistant for floating-point precision,” in International Conference on High Performance Computing, Networking, Storage and Analysis. ACM, 2013, p. 27.
[Silv16] D. Silver et al., “Mastering the game of Go with deep neural networks and tree search”, Nature, vol. 529, no. 7587, pp. 484–489, Jan. 2016.
[SqueezeNet] Iandola, F. N., Moskewicz, M. W., Ashraf, K., Han, S., Dally, W. J., and Keutzer, K. Squeezenet: Alexnet-leve accuracy with 50x fewer parameters and <1mb model size. CoRR, abs/1602.07360, 2016.
[Sung15] W. Sung, S. Shin, and K. Hwang, “Resiliency of Deep Neural Networks under Quantization,” CoRR, vol. abs/1511.06488, 2015.
[Sze17] V. Sze, Y. H. Chen, T. J. Yang and J. S. Emer, “Efficient Processing of Deep Neural Networks: A Tutorial and Survey,” in Proceedings of the IEEE, vol. 105, no. 12, pp. 2295-2329, Dec. 2017.
[Tann17] H. Tann, S. Hashemi, R. I. Bahar and S. Reda, “Hardware-software codesign of accurate, multiplier-free Deep Neural Networks,” 54th ACM/IEEE Design Automation Conference (DAC), 2017, pp. 1-6.
[Zhi17] L. Zhiqiang, Y. Dou, J. Jiang, J. Xu, S. Li, Y. Zhou, and Y. Xu, “Throughput- Optimized FPGA Accelerator for Deep Conv. Neural Networks”, ACM Trans. on Reconfig. Tech. and Syst., vol. 10, no. 3, 2017, p 17.