Charbel Sakr

PhD Candidate in ECE - University of Illinois at Urbana-Champaign
Phone Number: (217)6932866 / E-mail Address: sakr2@illinois.edu / Website: sakr2.web.engr.illinois.edu
Address: Office #411, Coordinated Science Laboratory, 1308 W Main St, Urbana, IL 61801

Education

- **University of Illinois at Urbana-Champaign** 2015-Present Illinois, USA
  - **PhD Candidate** in **Electrical and Computer Engineering** since May 2017.
  - **Masters of Science** in **Electrical and Computer Engineering** 2015-2017.

    Thesis title: *Analytical Guarantees for Reduced Precision Fixed-Point Margin Hyperplane Classifiers.*

    Thesis Supervisor: *Prof. N. Shanbhag.*

    Graduating GPA: **4.0**/4.
- **American University of Beirut** 2011-2015 Lebanon
  - **Bachelor** in **Engineering** (BE) in **Computer and Communications Engineering**

    Graduating GPA: **93.47**/100 (Equivalent to **4.0**/4).

    Recipient of a **Minor Mathematics** and a graduation **High Distinction**.

Research Experience

- **University of Illinois at Urbana-Champaign**
  - **Graduate Research Assistant** since the **Fall term 2015**, under the supervision of **Dr. Naresh Shanbhag**.
  - **Research interests**: I am broadly interested in **low-complexity**, **resource constrained Machine Learning** and **Signal processing**. A strong component of my research is in the understanding of **precision vs. accuracy trade-offs** in fixed-point learning systems, particularly, **deep neural networks**. I am also interested in low-cost *implementation* and *training* of neural networks, *statistical error compensation*, and *efficient memory systems*.
  - **Research Projects**

    Fixed-Point Neural Networks: This work leverages the theory developed in my earlier work on fixed-point margin hyperplane classifiers (listed below). We showed that by leveraging the back-propagation algorithm, we can understand the trade-offs between accuracy and precision of fixed-point neural networks. I am the main contributor of this project. One paper on the topic has been published at the 2017 *International Conference on Machine Learning* (**ICML'17** [1]). A follow up work on this topic with a more fine-grained analysis and improved empirical results was published at the 2018 *International Conference on Acoustics, Speech, and Signal Processing* (**ICASSP'18** [2]). Applications of the proposed theory to deep learning with biomedical datasets were published at the 2018 *Biomedical Circuits and Systems Conference* (**BioCAS** [3]).

    Per-Tensor Fixed-Point Quantization of the Back-Propagation Algorithm: Many works (including our own) have focused on reducing the complexity of neural networks at inference time via precision optimization. However, the much harder problem of reduced

precision training remains largely unresolved. In this work, we analyzed and determined precision requirements for training neural networks when all tensors, including back-propagated signals and weight accumulators, are quantized to fixed-point format. I am the main contributor of this project which culminates in a published paper at the 2019 *International Conference on Learning Representations* (**ICLR 2019** [4]).

Fixed-Point Hyperplane Margin Classifiers: This research seeks to bring rigor to the design of fixed-point learning systems which is currently being done using trial and error. Specifically we characterized the precision to accuracy trade-off of support vector machines (SVM) and general margin hyperplane classifiers. I am the lead contributor in this project. One paper on the topic has been published at the 2017 *International Conference on Acoustics, Speech, and Signal Processing* (**ICASSP'17** [5]). An extension of this work with generalized results for hyperplane classifiers with non-linear input and output maps, as well as improved empirical results form the basis of my *Masters Thesis* as well as a a paper in the June 2019 *Journal on Emerging and Selected Topics in Circuits and Systems* (**JETCAS'19** [6]).

Fundamental Limits on the Precision of In-Memory Architectures: This research seeks to obtain a fundamental understanding on the computational limits of in-memory architectures. A precision/SNR analysis is performed for different in-memory compute models and architectures. The fundamental aspect of this work makes its contribution highly theoretical while its application to in-memory architectures renders it highly practical. I am the main co-contributor of this project. Our work has already appeared as in a published paper at the 2020 *IEEE/ACM International Conference on Computer Aided Design* (**ICCAD'20** [7]). A more extensive and detail version of our research is currently being prepared for publication as a forthcoming journal paper.

KeyRam: This collaborative project implements a solid state circuits mapping a recurrent attention model (RAM). The RAM comprises of several layers of computations inside a feedback loop. In-memory computing is used to map the largest layers in the model which have lenient precision requirements. The smaller layers, which are more stringent are implemented using a novel digital matrix-vector-multiply (MVM) unit which I designed and implemented. Our chip is described in a published paper at the 2020 *IEEE Custom Integrated Circuits Conference* (**CICC'20** [8]) as well as a a paper in the 2020 *Journal on Solid State Circuits* (**JSSC'20** [9]).

PredictiveNet: This project proposes a simple but highly efficient architectural idea to reduce the computational cost of Convolutional Neural Networks (CNN). The idea is to decompose the arithmetic computation and predict sparse outputs efficiently. My contribution in this project was an analytical validation of the technique. One paper on the topic has been published at the 2017 *International Symposium on Circuits and Systems* (**ISCAS'17** [10]).

Compute-Sensor: This project aims to bring computation to the bitlines and cross-bitlines of a sensory array using mixed-signal techniques. My contribution was setting up the algorithm and validation dataset as well as post layout verifications.

- **IBM T. J. Watson Research Center**
  - **Research Intern** during both **Summer & Fall 2018** in the **Accelerator Architectures and Machine Learning** group.

    I worked under the supervision of **Dr. Kailash Gopalakrishnan**. The internship focused on the two topics of reduced precision training of deep neural networks as well as distributed learning algorithms with high parallelism and low communication costs. In particular, under the scope of deep learning with reduced precision floating-point arithmetic,

one breakthrough was achieved. We developed a theoretical framework able to predict accumulation bit-width (in the mantissa sense) requirements for all three deep learning GEMMs. This work was published at the 2019 *International Conference on Learning Representations* (**ICLR 2019** [11]).

- **Research Intern** during the **Summer 2017** in the **Accelerator Architectures and Machine Learning** group.

  I worked under the supervision of **Dr. Kailash Gopalakrishnan**. The internship focused on the topic of training deep neural networks with reduced numerical precision. Specifically, I worked on a method to use true gradient based learning for binary activated network. Part of this work was published at the 2018 *International Conference on Acoustics, Speech, and Signal Processing* (**ICASSP'18** [12]).

- **University of Toronto**

  - **Research Intern** during the **Summer term 2014** under the supervision of **Dr. Farid Najm**.

    The research was on Computer-Aided Design (CAD) for integrated circuits. My work involved the development of techniques for macromodeling parts of an on-die power grid in order to ensure the safety on its internal nodes.

Professional Service

- **Conference Organization**: I have served in conference organizing committees as follows:
  - **The 2019 14^{th} CSL Student Conference (CSLSC'19)**: *General Chair* - my responsibilities included preparing and running the whole 3-day conference including: forming an organizing committee, scheduling the conference, assigning conference sessions, securing sponsorships, putting together a job fair, and other conference management related activities.
  - **The 2018 13^{th} CSL Student Conference (CSLSC'18)**: *Session Chair* - I chaired one of the technical sessions of the conference titled 'Information Processing in Silicon', my responsibility included selecting and inviting a keynote speaker, selecting student talks, running the session, securing sponsorship, and other session management related activities.
  - **The 2017 12^{th} CSL Student Conference (CSLSC'17)**: *Committee member*.
- **Paper Reviewing**: I have served as a reviewer for the following publication venues:
  - **The 2020 International Conference on Learning Representations (ICLR'20)**.
  - **The 2021 International Conference on Learning Representations (ICLR'21)**.
  - **The 2020 Conference on Neural Information Processing Systems (NeurIPS'20)**.
  - **The 2019 Conference on Neural Information Processing Systems (NeurIPS'19)**.
  - **The 2020 International Conference on Machine Learning (ICML'20)**.
  - **The 2019 IEEE Journal on Emergin and Selected Topics in Circuits and Systems (JETCAS'19)**.
  - **The 2020 SIAM Journal on Mathematics of Data Science (SIMODS'20)**.
  - **The 2020 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS'20)**.
  - **The 2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS'19)**.
  - **The 2020 NeurIPS Reproducibility Challenge**.

Awards

- **Best in Session Award** at **Techcon** 2017.
- **ECE Rambus Fellowship in Electrical and Computer Engineering** for 2018-2019.
- **ECE Rambus Fellowship in Electrical and Computer Engineering** for 2019-2020.

Graduate Coursework

- **Digital IC Design**: Fall 2015 with *Prof. N. Shanbhag*: **A+**.
- **Graduate Level Digital Signal Processing**: Fall 2015 with *Prof. M. Do*: **A+**.
- **Analog IC Design**: Spring 2016 with *Prof. P. Hanumolu*: **A**.
- **Random Processes**: Spring 2016 with *Prof. V. Veeravalli*: **A+**.
- **Computational Inference and Learning**: Fall 2016 with *Prof. P. Moulin*: **A+**.
- **Machine Learning in Silicon**: Fall 2016 with *Prof. N. Shanbhag*: **A**.
- **Computational Complexity**: Spring 2017 with *Prof. A. Kolla*: **A**.
- **Statistical Learning Theory**: Spring 2017 with *Prof. B. Hajek*: **A**.
- **Learning Algorithms and Models**: Fall 2017 with *Prof. P. Viswanath*: **A**.
- **Computer Systems Organization**: Fall 2017 with *Prof. J. Torrellas*: **A**.
- **Integer Programming**: Spring 2018 with *Prof. S.R. Etesami*: **A**.
- **Inverse Problems and Learning**: Spring 2019 with *Prof. I Dokmanic*: **A-**.
- **Space, Time, and Matter**: Spring 2019 with *Prof. P. Phillips*: **A**.

Skills

- **Languages**: English, French, and Arabic. All three spoken and written fluently.
- **Computer**: I mostly code in Python and am proficient in several deep learning frameworks such as pyTorch and Theano. I also have experience in Matlab, C++, Verilog, Java, Haskell, Prolog, VHDL, Cadence. My all time favorite computer tool is LaTeX. Some of my codes can be found on my github profile at `github.com/charbel-sakr`.

Publications

[1] C. Sakr, Y. Kim, and N. Shanbhag, "Analytical Guarantees on Numerical Precision of Deep Neural Networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 3007–3016, 2017.

[2] C. Sakr and N. Shanbhag, "An Analytical Method to Determine Minimum Per-layer Precision of Deep Neural Networks," *International Conference on Accoustic, Speech, and Signal Processing (ICASSP)*, 2018.

[3] C. Sakr and N. Shanbhag, "Minimum precision requirements for deep learning with biomedical datasets," in *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2018.

[4] C. Sakr and N. Shanbhag, "Per-tensor fixed-point quantization of the back-propagation algorithm," in *International Conference on Learning Representations (ICLR)*, 2019.

[5] C. Sakr, A. Patil, S. Zhang, Y. Kim, and N. Shanbhag, "Minimum Precision Requirements for the SVM-SGD Learning Algorithm," *International Conference on Accoustic, Speech, and Signal Processing (ICASSP)*, 2017.

[6] C. Sakr, Y. Kim, and N. Shanbhag, "Minimum precision requirements of general margin hyperplane classifiers," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2019.

[7] S. K. Gonugondla, C. Sakr, H. Dbouk, and N. R. Shanbhag, "Fundamental limits on the precision of in-memory architectures," in *2020 International Conference on Computer Aided Design (ICCAD)*, IEEE/ACM, 2020.

[8] H. Dbouk, S. K. Gonugondla, C. Sakr, and N. R. Shanbhag, "Keyram: A 0.34 uj/decision 18 k decisions/s recurrent attention in-memory processor for keyword spotting," in *2020 IEEE Custom Integrated Circuits Conference (CICC)*, pp. 1–4, IEEE, 2020.

[9] H. Dbouk, S. K. Gonugondla, C. Sakr, and N. R. Shanbhag, "A 0.44 uj/dec, 39.9 us/dec, recurrent attention in-memory processor for keyword spotting," in *IEEE Journal on Solid State Circuits*, IEEE, 2020.

[10] Y. Lin, C. Sakr, Y. Kim, and N. Shanbhag, "PredictiveNet: An energy-efficient convolutional neural network via zero prediction," in *Circuits and Systems (ISCAS), 2017 IEEE International Symposium on*, pp. 1–4, IEEE, 2017.

[11] C. Sakr, N. Wang, C.-Y. Chen, J. Choi, A. Agrawal, N. Shanbhag, and K. Gopalakrishnan, "Accumulation bit-width scaling for ultra-low precision training of deep networks," in *International Conference on Learning Representations (ICLR)*, 2019.

[12] C. Sakr, J. Choi, Z. Wang, K. Gopalakrishnan, and N. Shanbhag, "True Gradient-based Training of Deep Binary Activated Neural Networks via Continuous Binarization," *International Conference on Accoustic, Speech, and Signal Processing (ICASSP)*, 2018.