Omar Mohamed Awad
Senior AI Researcher

1101 - 95 Oneida Crescent, Richmond Hill, L4B 0H5 ON +1 (437) 985-7774
omar.mo.awad@outlook.com linkedin.com/in/awadomar18
github.com/omarawad2 omarawad2.github.io google scholar

Professional Experience

October 2022 – Present — Senior AI Researcher, Huawei Technologies R&D — Toronto, Ontario

- Working on data-efficient training of large transformer-based models (e.g., BERT, RoBERTa, Conformer, T5, GPT-3, and ViT).
- Lead collaborations with university professors.
- Designed a novel dataset sampling algorithm for efficiently training DL models (patent in progress).
- Winner of the Data Application Acceleration Lab's Individual Award of Q1, 2023.
- Manager : Yang Liu

Data-efficient Training | Natural Language Processing | Computer Vision | PyTorch | TensorFlow | Perf. Optimization

August 2021 – October 2022 — Member of Technical Staff - Performance, Cerebras Systems — Toronto, Ontario

- Analysis/debugging/tuning of end-to-end performance (starting from model implementation in TensorFlow/PyTorch all down to microcode running on chip) of deep learning models on the Cerebras CS-2 Wafer.
- Performance modeling/projection of upcoming models (eg. Vision Transformer, Linformer) and kernels (e.g. attention) to be supported.
- Manager : Michael James, Mandeep Singh

Deep Learning | Performance Modeling | Kernel Optimization | Compilers

August 2020 – August 2021 — Machine Learning Research Engineer, Huawei Technologies R&D — Toronto, Ontario

- Optimize the training performance of various state-of-the-art NLP models (BERT, CPM, GPT-2/3) on Huawei's Ascend910 AI training server.
- Kernels development and performance optimization for Huawei's Ascend910 AI training server.
- Researching model compression techniques, e.g., low-rank tensor decomposition and layer truncation.
- Researching knowledge distillation techniques to improve accuracy of compressed models.
- Winner of the "Hardware Aware Efficient Training" competition at ICLR 2021. [Link]
- Managers : Yang Liu, Gordon Deng

PyTorch | Natural Language Processing | Computer Vision | Model Compression | Knowledge Distillation | Docker

September 2018 – July 2020 — Graduate Research Assistant, University of Toronto — Toronto, Ontario

- Design of a neural network training accelerator based on a novel processing element architecture that exploits fine-grain unstructured sparsity to increase the performance and energy efficiency of the training process by $1.47\times$ and $1.39\times$, respectively on average over the studied models.
- Development of a custom cycle-accurate trace-based simulator (C/C++) to model the execution time and memory access of the proposed accelerator compared to a baseline value-agnostic accelerator.
- Exploiting the narrow floating-point value distribution during training through exponent base-delta encoding compression to save off-chip memory bandwidth by 30% on average.
- Advisor : Prof. Andreas Moshovos.

C/C++ | PyTorch | Machine Learning Accelerator | Performance Modeling | RTL | Hardware Design | Computer Architecture

June 2018 – August 2018 — Research Intern, Opto-Nano-Electronics Lab, Cairo University — Cairo, Egypt

Conducted a comparative study on the performance-accuracy trade-offs (VHDL and Python) of using approximate multipliers such as Mitchell, Booth Radix-8, and Compressor in Convolution Neural Networks inference accelerators.

VHDL | Python | Approximate Computing | Machine Learning | ASIC

February 2017 – September 2017

Research Intern, Chair of Embedded Security, Ruhr University Bochum — Bochum, Germany

- Design of a novel zero-gate overhead hardware Trojan that is hard to detect using standard visual inspection hardware reverse engineering. The Trojan is based on the capacitive crosstalk effect between the chip interconnect.
- Trojan implemented (VHDL) in two VLSI designs : Advanced Encryption Standard (AES), and the OR1200 processor to controllably leak the encryption key, and trigger privilege escalation, respectively. Both chips were designed using NanGate FreePDK45 library without violating Design Rule Check (DRC).

VHDL | Verilog | Hardware Security | ASIC | Hardware Trojans | VLSI

August 2016 – September 2016

Research Intern, Opto-Nano-Electronics Lab, Cairo University — Cairo, Egypt

Hardware acceleration (VHDL) of AES, and RC6 encryption algorithms on Xilinx Spartan-3E FPGA board.

VHDL | Cryptography | FPGA | Hardware Acceleration

Education

September 2018 – July 2020

M.A.Sc., Electrical and Computer Engineering, University of Toronto — Toronto, Ontario

- **Thesis Topic :** Exploiting Fine-Grain Sparsity to Accelerate Neural Network Training. **GPA :** 4.0/4.0.
- **Supervisor :** Prof. Andreas Moshovos.

Machine Learning | Hardware Acceleration | Computer Architecture Exploration | Performance Modeling

September 2013 – July 2018

B.Sc., Electrical and Electronics Engineering, German University in Cairo — Cairo, Egypt

- **Thesis Topic :** Implementation of Hardware Trojans in ASIC Chips based on Routing Capacitive Crosstalk.**GPA :** 3.87/4.0.
- **Supervisor :** Prof. Christof Paar - Ruhr University Bochum, Germany.

Hardware Security | ASIC | Physical Design

Software Skills

Programming	**C, C++, Python, CUDA**, Matlab
Scripting	Perl, TCL, Bash
ML Models	CNN, RNN, Transformer-based (BERT, GPT-2/3, CPM, Vision Transformer)
ML Frameworks	PyTorch
DevOps Tools	Git, Docker, JIRA

Hardware Skills

Design Tools	Intel Quartus Prime, Xilinx ISE, Synopsys Design Compiler, HSPICE & HSIM,Cadence SoC Encounter, Innovus & Virtuoso
Simulation Tools	ModelSim, VCS
HDL	Verilog, VHDL
Arch. Simulators	SimpleScalar (modification), DNNsim (development)
Memory Compiler	CACTI

Scholarships & Awards

May 2023	**Winner of the Data Application Acceleration Lab's Individual Award of Q1, 2023**
May 2021	**Winner of the "Hardware Aware Efficient Training" competition at ICLR 2021. [Link]**
October 2020	**Winner of Huawei Quarterly Outstanding Contribution to Project Award**
2019, 2020	**University of Toronto Edward S. Rogers Sr. Graduate Scholarship for 2 years**
2017	**Ruhr University Bochum Undergraduate Research Award for 1 year**
2013-2018	**German University in Cairo High School Excellence Scholarship for 5 years**

Publications

2023 M. Elgammal, **O. Mohamed Awad**, I. Edo, A. Moshovos, V. Betz, "cuSCNN : an Efficient CUDA Implementation of Sparse CNNs", HEART '23 : Proceedings of the 13th International Symposium on Highly Efficient Accelerators and Reconfigurable Technologies [Link]

2021 **O. Mohamed Awad**, M. Mahmoud, I. Edo, A. Hadi Zadeh, C. Bannon, A. Moshovos, "PRaker : A Processing Element for Accelerating Neural Network Training", 54th IEEE/ACM International Symposium on Microarchitecture (MICRO), 2021. **[Acceptance Rate : 21%]** [PDF][Lightning][Main Talk]

2021 **O. Mohamed Awad**, H. Hajimolahoseini, M. Lim, G. Gosal, W. Ahmed, Y. Liu , G. Deng, "Improving ResNet-9 Generalization Trained on Small Datasets", Hardware Aware Efficient Training (HAET) at ICLR 2021. [PDF]

2020 A. Hadi Zadeh, I. Edo, **O. Mohamed Awad**, A. Moshovos, "GOBO : Quantizing Attention-Based NLP Models for Low Latency and Energy Efficient Inference", 53rd IEEE/ACM International Symposium on Microarchitecture (MICRO), 2020. **[Acceptance Rate : 19%]** [PDF]

2020 M. Mahmoud, I. Edo, A. Hadi Zadeh, **O. Mohamed Awad**, J. Albericio, A. Moshovos, "TensorDash : Exploiting Sparsity to Accelerate Neural Network Training", 53rd IEEE/ACM International Symposium on Microarchitecture (MICRO), 2020. **[Acceptance Rate : 19%]** [PDF]

2019 A. Delmás, S. Sharify, I. Edo, D. Malone Stuart, **O. Mohamed Awad**, P. Judd, M. Mahmoud, M. Nikolic, K. Siu, Z. Poulos, and A. Moshovos, "ShapeShifter : Enabling Fine-Grain Data Width Adaptation in Deep Learning", 52nd IEEE/ACM International Symposium on Microarchitecture (MICRO), 2019. **[Acceptance Rate : 23%]** [PDF]

2019 C. Kison, **O. Mohamed Awad**, M. Fyrbiak, C. Paar, "Security Implications of Intentional Capacitive Crosstalk", IEEE Transactions on Information Forensics and Security, 2019. [PDF]

Patents

2020 **O. Mohamed Awad**, M. Mahmoud, and A. Moshovos, "System and method for accelerating training of deep learning networks", **U.S. 63/054,502 (2020).**

2020 A. Hadi Zadeh, I. Edo, **O. Mohamed Awad**, and A. Moshovos, "Quantization for neural network computation", **U.S. 17/130,690 (2020).**

2020 A. Hadi Zadeh, I. Edo, **O. Mohamed Awad**, and A. Moshovos, "GOBO : Quantizing attention-based NLP models for low latency and energy efficient inference", **U.S. 63/082,009 (2020).**

Selected Courses

Grad Courses	Parallel Computer Architecture and Programming (A+), Reconfigurable Computing and FPGA Architecture (A+), Introduction to Machine Learning (A), Programming Massively Parallel Microprocessors (A), Advanced Computer Architecture (A+)
Undergrad Courses	System-On-a-Chip (A+), Advanced Microelectronics Lab (A+), Programmable Logic Circuits (A+), Very Large Scale Integration (A+), Micro-Computer Applications (A+).
Online Courses	Computer Architecture (Princeton University), Neural Networks and Deep Learning (deeplearning.ai).

Selected Projects

Compressed-memory sparse DNN inference accelerator on GPU — 2019

github.com/Omar-Awad/SCNN_GPU2

Compressed-memory sparse DNN inference accelerator on NVIDIA GeForce GTX980, achieving a speedup up to $115\times$ and $170\times$ on image classification and computational imaging models, respectively compared to an efficient openMP multi-threaded CPU implementation.

Machine Learning | CUDA | openMP | Computer Vision

Logical-to-physical RAM mapping for FPGAs — 2019

Logical-to-physical RAM mapping CAD tool for FPGAs with various types of physical RAMs.

CAD | FPGA | C/C++

5-stages pipelined MIPS processor — 2016

github.com/Omar-Awad/encrypted-MIPS

5-stages pipelined MIPS processor with 10-instructions and encrypted memory using Xilinx ISE.

MIPS | VHDL | FPGA

References

Available Upon Request