

# Omar MOHAMED AWAD

## Senior ML Research Engineer

📍 Mississauga, Ontario, Canada 📞 +1 (437) 985-7774

@ omar.mo.awad@outlook.com in linkedin.com/in/awadomar18  
github.com/omarawad2 omarawad2.github.io google scholar

### PROFESSIONAL EXPERIENCE

October 2022  
Present



#### Senior ML Research Engineer, HUAWEI TECHNOLOGIES R&D

📍 Toronto, Ontario

- > Mentoring new hires in the team during their 3-month ramp-up program (setting goals, technical/non-technical mentorship, review/evaluate progress)
- > Interviewing candidates for our internal SW/HW co-design project (exploring the interplay between ML arch/training algos with HW to influence next release of Huawei's custom ML accelerator).
- > Designing efficient Transformer + SSM Hybrid LLMs.
- > Leading a HW/SW-codesign project on sub-8-bits training and inference of LLMs.
- > Training Memory-augmented LLMs to support long/unlimited context.
- > Researching data-efficient training algorithms that provide significant time-to-accuracy savings in pretraining LLMs.
- > Working on a data-efficient training library that provides various out-of-the-box algorithms to accelerate the training of any arbitrary LLM.
- > Main contributor of 2 internal patents (one for a novel dataset sampling method to accelerate model training, and another novel method to accelerate MoE transformer models).
- > Winner of the Data Application Acceleration Lab's Individual Award of Q1, 2023. [\[Link\]](#)
- > Manager : [Yang Liu](#)

Generative AI Large Language Models NLP Data-efficient Training Fintuning Inference PyTorch  
Perf. Optimization Team Leadership Deepspeed Megatron-LM LLaMa-Factory Accelerate

September 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 1.47x and 1.39x, 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 | Research Intern, **OPTO-NANO-ELECTRONICS LAB, CAIRO UNIVERSITY** 📍 Cairo, Egypt  
 August 2018 | 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 | Research Intern, **CHAIR OF EMBEDDED SECURITY, RUHR UNIVERSITY BOCHUM** 📍 Bochum, Germany  
 September 2017 |  **RUB**  
 > 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

## EDUCATION

September 2018 | M.A.Sc., Electrical and Computer Engineering, **UNIVERSITY OF TORONTO** 📍 Toronto, Ontario  
 July 2020 |   
 > **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 | B.Sc., Electrical and Electronics Engineering, **GERMAN UNIVERSITY IN CAIRO** 📍 Cairo, Egypt  
 July 2018 |   
 > **Thesis Topic** : Implementation of Hardware Trojans in ASIC Chips based on Routing Capacitive Crosstalk. **GPA** : 3.87/4.0 (Excellent with Highest Honors, top 1%).  
 > **Supervisor** : [Prof. Christof Paar](#) - Ruhr University Bochum, Germany.  
 Hardware Security ASIC Physical Design

## SOFTWARE SKILLS

<b>Programming</b>	Python, C, C++, CUDA, Java
<b>Scripting</b>	Perl, TCL, Bash
<b>ML Models</b>	Convolutional Neural Networks, Recurrent Neural Networks, Transformers, State Space Models, Hybrid Architectures
<b>ML Frameworks</b>	PyTorch, TensorFlow
<b>ML Acceleration Libraries</b>	Megatron-LM, Llama-factory, Triton, Accelerate, TensorRT, Deepspeed
<b>DevOps Tools</b>	Git, Docker, Conda, JIRA

## HARDWARE SKILLS

<b>HDL</b>	Verilog, VHDL
<b>Arch. Simulators</b>	SimpleScalar (modification), <a href="#">DNNsim</a> (development)
<b>Memory Compiler</b>	CACTI

## SCHOLARSHIPS & AWARDS

May 2023	Winner of the Data Application Acceleration Lab's Individual Award of Q1, 2023. <a href="#">[Link]</a>
May 2021	Winner of the "Hardware Aware Efficient Training" competition at ICLR 2021. <a href="#">[Link]</a>
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

## PATENTS

2023	<b>O. Mohamed Awad</b> , M. Mahmoud, and A. Moshovos, "System and method for accelerating training of deep learning networks", <b>US20230297337A</b> . <a href="#">[Link]</a>
2022	A. Hadi Zadeh, I. Edo, <b>O. Mohamed Awad</b> , and A. Moshovos, "Quantization for neural network computation", <b>US20220092382A1</b> <a href="#">[Link]</a>

## PUBLICATIONS

- 2024 M. Nikolić, G. Hacene, C. Bannon, A. Lascorz, M. Courbariaux, **O. Mohamed Awad**, I. Vivancos, Y. Bengio, V. Gripon, A. Moshovos, “Bitpruning : Learning bitlengths for aggressive and accurate quantization”, 2024 IEEE International Symposium on Circuits and Systems (ISCAS). [\[Link\]](#)
- 2024 F. Ataiefard, W. Ahmed, H. Hajimolahoseini, S. Asani, F. Javadi, M. Hassanpour, **O. Mohamed Awad**, A. Wen, K. Liu, Y. Liu, “SkipViT : Speeding Up Vision Transformers with a Token-Level Skip Connection”, Association for the Advancement of Artificial Intelligence (AAAI 2024). [\[Link\]](#)
- 2023 F. Javadi, W. Ahmed, H. Hajimolahoseini, F. Ataiefard, M. Hassanpour, S. Asani, A. Wen, **O. Mohamed Awad**, K. Liu, Y. Liu, “GQKVA : Efficient Pre-training of Transformers by Grouping Queries, Keys, and Values”, 37th Conference on Neural Information Processing Systems (NeurIPS 2023). [\[Link\]](#)
- 2023 H. Hajimolahoseini, **O. Mohamed Awad**, W. Ahmed, A. Wen, S. Asani, M. Hassanpour, F. Javadi, M. Ahmadi, F. Ataiefard, K. Liu, Y. Liu, “SwiftLearn : A Data-Efficient Training Method of Deep Learning Models using Importance Sampling”, 37th Conference on Neural Information Processing Systems (NeurIPS 2023). [\[Link\]](#)
- 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 H. Hajimolahoseini, M. Rezagholizadeh, V. Partovinia, M. Tahaei, **O. Mohamed Awad**, Y. Liu, “Compressing Pre-trained Language Models using Progressive Low Rank Decomposition”, 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks · Dec 6, 2021. [\[PDF\]](#)
- 2021 **O. Mohamed Awad**, M. Mahmoud, I. Edo, A. Hadi Zadeh, C. Bannon, A. Moshovos, “FPRaker : 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. M. 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. [\[PDF\]](#)

## 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).