Farshid Varno

Intro

Seasoned AI and hardware co-design engineer with over 15 years of combined academic and industry experience spanning hardware design, software development, and AI research. Expertise includes designing and simulating hardware modules, developing custom AI kernels and non-linear approximators, and pioneering innovations in Federated and Transfer Learning. A proven track record of impactful contributions, including leading patented advancements in AI accelerator architectures and Transfer Learning, developing FPGA-based telecommunication modules, and optimizing hardware-software integration for AI systems. Passionate about bridging AI, software, and hardware to develop cutting-edge solutions for next-generation computing challenges, while actively exploring opportunities to lead and drive innovation through technical management.

Experience

Research & Industry

Research Scientist, Rain

August 2023–present, San Francisco, USA

- Leading Prototype of LUT-based non-linear fns.
- Led the design of a sparse VMM unit, and an efficient online FP Softmax unit with Base-2 conversion.
- Led on-device training and domain adaptation for resource-constrained environments.
- Principal inventor on three patents for Softmax HW and CIM architecture innovations.
- Conducting AI/HW co-optimization, focusing on mixed-precision quantization, RISC-V extensions, and PyTorch Dynamo integration.
- Designing and simulating HW units for CIM tech. using SystemC, QEMU, and AI performance modeling to optimize LLMs and attention architectures.

Research Scientist, Imagia

May 2018–March 2022 (internship period included), Montreal, Canada

- Led research on Federated Learning optimization, achieving SoTA performance via drift elimination.
- Researched Transfer Learning, out-of-distribution generalization, Meta-Learning, and Few-Shot Learning.
- Helped design an AI library to enhance Imagia's research efforts.
- Published a patent on Transfer Learning as the lead inventor.
- Contributed to porting Polyaxon to DGX clusters, enhancing resource efficiency for AI workloads.

Research Assistant, Institute for Big Data Analytics

May 2017–May 2018, Halifax, Canada

- Collaborated with Harvard University to research human behavior prediction from fMRI data using advanced modeling techniques.
- Developed a CNN framework for detecting aircraft corrosion with D-Sight technology (DAIS).
- Designed a new CUDA implementation for minimum distance calculations in AIS-GIS streaming data, achieving a substantial performance improvement.
- Researched sparsity, activation functions, and normalization for efficient machine learning models.

Data Scientist (part-time), Cognitive Health and Recovery Research Lab

Mar 2020–Jun 2020, Halifax, Canada

- Integrated and visualized clinical data to support cognitive health research.
- Investigated post-operative cognitive dysfunction in elderly patients through data analysis.
- Analyzed surgical time series data (e.g., anesthesia depth, patient vitals) to identify patterns and insights.

FPGA Engineer, Kara Telephone Co.

Jun 2013–Jun 2014, Tehran, Iran

- Designed, implemented, and integrated TDM switches on FPGAs, supporting up to 16k x 16k channels.
- Developed a multi-channel I2C master controller for 16 modules with error checking and correction.
- Designed and implemented SPI and USART peripheral interfaces, ensuring seamless system integration.
- Worked with embedded processors and RTOS, optimizing hardware-software interaction.
- Led speed optimization efforts for FPGA designs on Altera Cyclone series.

RTL Designer Intern, SarvNet Telecommunication Inc.

Jul 2012–Sep 2012, Isfahan, Iran

- Designed and implemented AES modules for encryption in STM4 lines, ensuring efficient performance.
- Developed resource-sharing mechanisms to support both 128-key and 256-key AES modes, adapting dynamically based on the selected encryption mode.
- Area optimization for FPGA designs, targeting Xilinx Virtex 4 and 6 series to minimize resource usage.

Teaching

- Adjunct Professor, Computer Architecture, Chehelsotoon Institute for Higher Education, Fall 2015
- Adjunct Professor, System Programming, Chehelsotoon Institute for Higher Education, Fall 2015
- Co-instructor, Machine Learning for Big Data (CSCI-6515), Dalhousie University, Fall 2020
- Teaching Assistant, Machine Learning for Big Data (CSCI-6515), Dalhousie University, Fall 2018
- Teaching Assistant, Digital Circuits (ECED-2200), Dalhousie University, Winter 2016
- Teaching Assistant, System Analysis (ECED-3401), Dalhousie University, Fall 2016
- Teaching Assistant, Java Programming, University of Guilan, Winter 2009
- Teaching Assistant, Algorithms, University of Guilan, Winter 2010

Background

Education

- Ph.D., Computer Science. Dalhousie University. 2017–2023, CGPA: 4.19
- M.Sc., Computer Architecture. University of Isfahan. 2012–2015, CGPA: 4.02
- B.Sc., Comuter Engineering, Guilan University. 2008–2012.

Skills

- Programming & AI Frameworks: Python, C++, PyTorch, Triton, CUDA
- RTL Design & Simulation: VHDL, Verilog, SystemC
- AI Optimization & Deployment: On-device training, quantization & compression, OCP microscaling formats
- CI/CD, Build & MLOps Tools: GitHub Actions, Bazel, Poetry, Tox, Polyaxon, MLflow
- Markup & Documentation: LaTeX, Markdown, RestructuredText, Mermaid

Selected Publications

Papers

- Varno, Farshid, Marzie Saghayi, Laya Rafiee, Sharut Gupta, Stan Matwin, and Mohammad Havaei. "Minimizing Client Drift in Federated Learning via Adaptive Bias Estimation." *European Conference on Computer Vision. – **ECCV*** (2022).

- Varno, Farshid, Lucas May Petry, Lisa Di Jorio, and Stan Matwin. "Learn Faster and Forget Slower via Fast and Stable Task Adaptation." *arXiv preprint arXiv:2007.01388* (2020).

- Varno, Farshid, Behrouz Haji Soleimani, Marzie Saghayi, Lisa Di Jorio, and Stan Matwin. "Efficient neural task adaptation by maximum entropy initialization." *arXiv preprint arXiv:1905.10698* (2019).

- Jiang, Xiang, Mohammad Havaei, Farshid Varno, Gabriel Chartrand, Nicolas Chapados, and Stan Matwin. "Learning to learn with conditional class dependencies." In *international conference on learning representations. – **ICLR*** (2018).

- Saghayi, Marzie, Jonathan Greenberg, Christopher O'Grady, Farshid Varno, Muhammad Ali Hashmi, Bethany Bracken, Stan Matwin, Sara W. Lazar, and Javeria Ali Hashmi. "Brain network topology predicts participant adherence to mental training programs." ***Network Neuroscience*** 4, no. 3 (2020): 528-555.

Patent

- Varno, Farsheed, Behrouz Haji Soleimani, Marzie Saghayi, Lisa Di Jorio, and Stan Matwin. Method and system for initializing a neural network. https://patents.google.com/patent/WO2020225772A1. _ EP WO CA CN_ (2020)

- Three patents in provisioning stage (will update soon).

Recognition

Honors

- **Vice-president of Public Relations**, Toastmasters International, Dal Toastmasters, 2020.
- **Mitacs Accelerate Award**, 56k CAD, 2021–2022.
- **Scotia Scholar Award**, 45k CAD, Research Nova Scotia, 2019–2021.
- **Best Graduate Student Research Award**, Big Data Congress, Sep 2017.
- Selected **Conference Program Committee Member** for ICLR (2020), KDD (2017), and Confoo (2023).
- Recognized as a **Reviewer** for leading AI conferences: CVPR 2025 (1 paper), ECCV 2024 (6), CVPR 2024 (2), ICCV 2023 (3), CVPR 2023 (5), ECCV 2022 (2), FedVision 2023 (2).
- **1st Rank Student Recognition**, University of Isfahan, Mar 2015.

Leadership & Mentoring

- Mentored Bachelor's and Master's students through various programs and occasions, and college students specifically through AI4ALL (2024).
- Experienced leading teams of 2–3 researchers during multiple projects.