Farshid Varno

Intro

AI, software, and hardware are often seen as distinct fields in modern computing. However, building a winning AI system requires combining these elements rather than treating them separately. This is where I come in. My unique blend of skills and experience enables me to integrate these components seamlessly, ensuring they work together as a cohesive system.

In summary I have:

- 5+ years experience as AI Research Scientist
- 1 years experience as Data Scientist
- 2+ years experience as FPGA Engineer/RTL Designer

Experience

Research & Industry

Research Scientist, Rain

August 2023–present, San Francisco, USA

- Design and simulation of digital hardware units for Rain's next-generation Compute-In-Memory (CIM) technology, tailored to meet the demands of trending LLMs (e.g., LLaMA, Mistral).
- Leading the design and prototyping of novel architectures for FP LUT-based non-linear units, including SiLU, Swish, GLU, and Exp in Softmax, among other activation functions.
- Quantitative AI/HW optimization of LLM architectures, focusing on specific model requirements (layer-wise mixed-precision quantization, quantization block sizes) for models like LLaMA2/3/3.1, Mistral/Mixtral, Gemma, Phi2, and Whisper (audio model with attention mechanisms).
- Leading the design of an efficient online Softmax unit with Base-2 conversion and FP input/output, addressing known bottlenecks in throughput and memory consumption within the attention mechanism.
- Driving critical design decisions with the Architecture and Simulation teams on factors such as OCP-Microscaling FP formats, optimal LUT sizing, RISC-V extensions.
- Leading the development of a lightweight, high-frequency sparse vector-vector multiplication unit optimized for LLM workloads.
- Conducting R&D of novel machine learning algorithms for edge devices, including applications in Rain's current and future CIM products.
- Principal inventor on three patents—two for Softmax hardware and another for CIM architecture (all currently in the provisioning stage and confidential).

Research Scientist, Imagia

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

- Research on *Federated Learning* optimization led to SoTA performance via drift elimination
- Research on Transfer Learning led to a filed patent
- Research on multiple Meta Learning and Few-shot Learning projects
- Research on Multi-hypothesis Transfer Learning and out of distribution generalization
- Collaborated with R&D team in designing an AI library for Imagia research
- Collaborated with IT in porting Polyaxon on a cluster of NVIDIA DGX systems

Research Assistant, Institute for Big Data Analytics

May 2017–May 2018, Halifax, Canada

- Research on predicting human behaviour from *fMRI* data
- Developing a *CNN* framework for detecting corrosion in aircrafts using *D-Sight* technology (*DAIS*)
- Optimizing calculation of minimum distance to shore from *AIS-GIS* streaming data using *CUDA* and *OpenMP*
- Research on sparsity, activation functions and normalization

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

Mar 2020–Jun 2020, Halifax, Canada

- Clinical data integration and visualization
- Investigating post-operative cognitive dysfunction in elderly patients
- Analyzing surgical time series data (anesthesia depth, patients' vitals, ...)

FPGA Engineer, Kara Telephone Co.

Jun 2013–Jun 2014, Tehran, Iran

- Design & Imp. of TDM switches on FPGAs supporting up to 16k x 16k channels (in VHDL)
- Multi-channel I2C master controller supporting 16 modules with error checking & correction
- SPI & USART Peripheral interfaces
- Embedded Processors, RTOS
- Focus on speed optimization on Altera Cyclone series

RTL Designer Intern, SarvNet Telecommunication Inc.

Jul 2012–Sep 2012, Isfahan, Iran

- Design & Imp. of lightweight AES modules used in STM4 lines
- Multi-channel I2C master controller supporting 16 modules with error checking & correction.
- SPI & USART Peripheral interfaces.
- Focus on area optimization on Xilinx Virtex 4, 6 series

Teaching

- Co-instructor, ML for Big Data, CSCI-6515, Dalhousie University, Fall 2020
- Teacher Assistant, ML for Big Data, CSCI-6515, Dalhousie University, Fall 2018
- Teacher Assistant, Digital Circuits, ECED-2200, Dalhousie University, Winter 2016
- Teacher Assistant, System Analysis, ECED-3401, Dalhousie University, Fall 2016
- Instructor, Computer Architecture, Chehelsotoon Inst. for Higher Edu, Fall 2015
- Instructor, System Programming, Chehelsotoon Inst. for Higher Edu, Fall 2015
- Teacher Assistant, Java Programming, University of Guilan, Winter 2009
- Teacher 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 languages: **Python**, Java, C/C++, Bash
- Deep learning frameworks: **PyTorch**, Keras, Tensorflow
- CI/CD platform: **Github Actions**
- Packaging, build and test tools: Bazel, Poetry, Tox
- MLOps, automation & AI scaling systems: **Polyaxon**, **MLflow**
- Target specific on-device training, Quantization & Compression, OCP Microscaling formats
- Machine learning libraries: Pandas, Scikit-learn, Numpy, Scipy
- Digital Circuit Design and FPGA engineering: **VHDL**, **Verilog**
- RISC-V instrcutions and extensions
- Markup languages: LaTeX, Markdown, RestructuredText, Mermaid
- Project Management: Agile, Scrum, Kanban, Jira, YouTrack

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)

Honors

Leadership & Volunter Work

- Vice-president of Public Relations, Toastmasters International, Dal Toastmasters, 2020.
- Mentoring college students, **AI4ALL**, 2024.
- Experienced leading teams of 2-3 researchers during several projects.
- Mentored two masters students, currently one working as Senior Data Scientists in Canada and the other one as Data Scientist in Brazil.
- Reviewer at European Conference on Computer Vision (**ECCV 2024**, 6 papers).
- Reviewer at Computer Vision and Pattern Recognition (**CVPR 2024**, 2 papers).
- Reviewer at International Conference on Computer Vision (**ICCV 2023**, 3 papers).

- Reviewer at Computer Vision and Pattern Recognition (**CVPR 2023**, 5 papers).
- Reviewer at European Conference on Computer Vision (**ECCV 2022**, 2 papers).
- Reviewer at the 2nd FedVision Workshop (**FedVision 2023**, 2 papers).
- Selected conference program committee member & volunteering experience:
  - International Conference on Learning Representations (ICLR, Remote 2020)
  - SIGKDD Conference on Knowledge Discovery and Data Mining (KDD, Halifax 2017).
  - Confoo (Montreal, 2023).

Awards & Recognition

- Accelerate Award, 56k CAD, Mitacs, 2021-2022
- Scotia Scholar Award, 45k CAD, Research Nova Scotia, 2019-2021
- Best Graduate Student Research Award, Big Data Congress, Sep 2017
- Nova Scotia University Student Bursary, Government of Nova Scotia, 2020-2022
- FGS's alloc. for outstanding status, 2k CAD, Dalhousie University, Aug 2017
- 1st Rank Student Recognition, University of Isfahan, Mar 2015

This resume is compiled by Farshid Varno with LaTeXand is also availabel in HTML format at https://farshid.varnio.com/resume/farshid/index.html