Seyedhamidreza Mousavi

+46736620799 — seyedhamidreza.mousavi@mdu.se — hamidrezamousavi — hamidmousavi0.github.io — Västerås, Sweden

Summary

- Over 5 years of research experience in **efficient deep learning** and **model compression**, focusing on the design of **compact and robust neural networks** for **edge and embedded systems**.
- Strong expertise in **Neural Architecture Search (NAS)**, **pruning**, **quantization**, **fault-tolerant learning**, and **adversarial robustness**, with extensive hands-on development in **PyTorch**.
- Solid theoretical and practical background in **Transformers**, **Large Language Models (LLMs)**, **Vision-Language Models (VLMs)**, **RAG systems**, **agentic AI**, and **diffusion models**.
- Hands-on experience deploying and optimizing models on **embedded platforms**, including **FPGA (AMD Artix™ 7 FPGA)**, **NVIDIA GPUs (Jetson Nano, A100, H100)**, and **microcontrollers (STM32)**, using **TensorRT** and **TinyEngine**.
- Proficient in **distributed training** on **multi-GPU systems** and in building **end-to-end MLOps pipelines**, including experiment tracking, model and data versioning (**GIT and DVC**), and scalable deployment using **Google Vertex AI**.
- **Applied ML researcher and engineer** with strong **problem-solving** skills, experienced in **simplifying complex problems**, cross-disciplinary **collaboration**, and **teaching**.

Experience

Mälardalen University **Jan 2022 – Present**
PhD Researcher — Efficient and Robust Deep Learning *Västerås, Sweden*

- Lead research on **efficient and Robust** deep neural networks for resource-constrained and safety-critical systems.
- Developed a **one-shot NAS** framework using **knowledge distillation**, enabling robust and efficient subnetworks under adversarial and resource constraints.
- Propose a parameterized activation function improving **fault tolerance** to hardware-level bit-flip errors.
- Integrate **pruning** and **ternerization** in NAS to provide more efficient models.
- Deploy **TinyDL** model for person detection on **STM32 Microcontrollers** using **TinyEngine**.
- Mentor junior interns and other PhD students in our research group and support course instruction for Deep Learning and Embedded Systems.
- Develop open-source tools for robust and reliable DNNs (ProARD- Robust and efficient DNNs: ProARD and Reliable DNNs: reliable-relu-toolbox)

Simon Fraser University **Jan 2020 – Jun 2022**
Machine Learning Researcher — Reliability in FPGAs and hardware accelerators *British Columbia, Canada (Remote)*

- Contribute to the ICCAD 2020 paper "Aadam: A Fast, Accurate, and Versatile **Aging-Aware Cell Library** Delay Model Using **Feed-Forward Neural Networks**," focusing on reliability-aware neural modelling.
- Contribute to the research on fault injection and adversarial robustness, including "Stealthy Attack on **Algorithmic-Protected DNNs** via Smart Bit Flipping"

Skills

- **Deep Learning & AI:** Efficient AI, Model Compression, NAS, Pruning, Quantization, Fault Tolerance, Adversarial Robustness
- **Frameworks & Systems:** PyTorch, HuggingFace, LangGraph, Horovod (Multi-Gpus training), TensorRT.
- **Distributed & MLOps:**: Multi-GPU Training, Horovod, DDP, MLOps, Vertex AI, GCP, DVC, Docker.
- **Programming**: Python, C/C++, VHDL, Latex.
- **Embedded & Hardware:**: FPGA (Vivado), STM32, NVIDIA Jetson, VHDL.
- **Soft Skills**: Problem-solving, Complex Problem simplification, cross-disciplinary collaboration, teaching.

Research Projects

FASTER AI: Fully Autonomous Safety- and Time-critical Embedded Realization of Artificial Intelligence

- Deep Learning Researcher from Mälardalen University.
- Collaboration with **KTH Royal Institute of Technology**, **RISE Research**, Institutes of Sweden, **EmbeDL** AB, **Ericsson** AB and **Saab**.
- **Role:** Research on **Pruning, quantization and knowledge distillation** to make efficient deep learning models.

AutoDeep: Automatic Design of Safe, High-Performance and Compact DL Models for Autonomous Vehicles.

- Deep Learning Researcher from Mälardalen University
- Collaboration with **Volvo-CE** and **Zenseact AB**
- **Role:** Develop Robust Neural Architecture Search methods to automatically design the robust and efficient architecture.

Research Interest

- **Efficient LLMs and VLMs**: Provide small multi-modal language models for edge devices.
- **Argentic AI and Tool use**: Provide automated argentic workflows to make action-based decisions.
- **Efficient Diffusion Models**: Provide Fast image and video generation with diffusion models.
- **Reasoning**: Finetuning LLMs to solve complex problems.

Education

Mälardalen University **Jun 2022 – Present**
PhD student of Computer Science
Thesis: Efficient Design and Training of Compact and Robust Deep Neural Networks

Shahid Bahonar University **Nov 2012 – Nov 2019**
Bachelor and Master student of Computer Engineering
Bachelor Thesis: Design and Implementation of ARM7-TDMIS on Xilinx Spartan 6 FPGA
Master Thesis: Reliability and Security Analysis of Deep Learning Models

Teaching Experience

Designed and teach lectures and hands-on labs in the following courses and semeniars:

- **Deep Learning at Mälardalen University** : Fall 2022, 2023, 2024, and 2025.
- **Computer Architecture at Mälardalen University**: Spring 2022, 2023, and 2024.
- **Enhancing the Reliability of Deep Learning Models at TSS conferance 2025 at Tallinn, Estonia**
- **Generative AI Seminar (Diffusion Models) at Mälardalen University 2025**

Recent Publications

- ProARD: Progressive adversarial robustness distillation: provide wide range of robust students. **IJCNN 2025**
  - One-shot NAS that reduces the computational cost by 60× and improves accuracy and robustness by 13% and 14%.
- DASS: Differentiable Architecture Search for Sparse Neural Networks. **TECS 2023**
  - Increases the accuracy of the sparse version of MobileNet-v2 from 73.44% to 81.35% with a 3.87× faster inference time.
- Aadam: A fast, accurate, and versatile aging-aware cell library delay model using neural network. **ICCAD 2020**
  - Enables aging-aware time analysis for large-scale circuits in 5 to 12 seconds to analyze a circuit with 138.9K to 255.3K gates.
- TAS: Ternarized Neural Architecture Search for Resource-Constrained Edge Devices. **DATE 2022**
  - Delivers 2.64% higher accuracy and 2.8 ×memory saving over competing methods with the same bit-width resolution
- Stealthy attack on algorithmic-protected dnns via smart bit flipping. **ISQED 2022**
  - Robustness of protected DNNs can significantly decrease under adversarial attack with a small number of bit-flips into the memory.
- Proact: Progressive training for hybrid clipped activation function to enhance resilience of DNNs. **Preprint 2025**
  - Improves the resilience of DNNs, with enhancements of up to 6.4x in high bit error rates
- Efficient On-device Transfer Learning using Activation Memory Reduction. **FMEC 2024**
  - Reduces peak activation memory and total memory costs of MobileNetV2 by 65% and 59%, respectively, at the cost of 3% accuracy drop
- Contextual range-view projection for 3D LiDAR point clouds. **Submitted to ICASSP 2026**
  - Preserves more instance points during projection, achieving up to a 3.1% mIoU improvement
- Analysing robustness of tiny deep neural networks. **ADBIS 2023**
  - Expanding the width of the blocks in MobileNet-tiny can improve the natural and robust accuracy.