Ali Hadi Zadeh

EA306 – 4, Engineering Annex, University of Toronto, ON, M5S 1A4 - Canada
✉ hadizadeh-AT-outlook.com • hadiza.de • AliHadizadeh

Research Highlights

Experienced Ph.D. Candidate in Computer Engineering from the University of Toronto with a demonstrated history of research in **Machine Learning** and **Computer Architecture**. During my Ph.D., we developed *model compression* methods and tailored *AI hardware accelerators* to improve the performance and energy efficiency of Attention-Based NLP models without any modifications to the models or fine-tuning them.

Education

- **University of Toronto – Vector Institute** — **Toronto, Canada**
  Ph.D. in Computer Engineering, ECE Dept. – Postgraduate Affiliate, — *2018–Present*
  - **Thesis:** Fast and Energy Efficient Machine-Learning-Based Natural Language Processing,
  - **Advisor:** Andreas Moshovos, **GPA:** 4 / 4.
- **Sharif University of Technology** — **Tehran, Iran**
  M.Sc. in Electrical Engineering, EE Dept., — *2015 – 2017*
  - **Thesis:** Digital Real-Time Simulation of Electrical Machine for Hardware-in-the-loop Applications,
  - **Advisors:** Matin Hashemi, Mostafa Parniani, **GPA:** 18.73 / 20, Dissertation: Excellent.
- **Shahid Chamran University** — **Ahvaz, Iran**
  B.Sc. in Electrical Engineering, EE Dept., **GPA:** 19.06 / 20. — *2011 – 2015*

Work Experience

- **Research Assistant/Ph.D. Candidate** — **Toronto, Canada**
  University of Toronto — *Sep 2018 – Present*
  - My research focuses on accelerating Deep Learning models such as Transformers by developing hardware-aware model compression techniques. Our solutions include: I) Developing **post-training quantization** methods that compress trained models without any need for fine-tuning, access to datasets or training resources, and II) Designing **hardware accelerators** and memory compression/decompression engines that maximize the performance and energy efficiency for quantized models.
- **Research Associate – Research Assistant/M.Sc. Candidate** — **Tehran, Iran**
  Sharif University of Technology — *Sep 2015 – Sep 2018*
  - Design and FPGA implementation of real-time simulators for power systems, capable of solving a system of differential equations in a few nano seconds, leveraging the Sherman–Morrison method to gradually update the matrix inversion computation.

Awards & Honors

- **Postgraduate Affiliate at Vector Institute** – Canada's Largest AI community, acceptance rate of 8%.
- **Student Travel Grant**: ISCA'22, IISWC'19.
- **Nominated for the Microsoft Research PhD Fellowship** among PhD candidates at UofT, 2021.
- **Top 5% Among M.S. Graduate Students**, EE Dept., Sharif University of Technology.
- **Nominated for Best Dissertation Award**, Sharif University of Technology among 200 M.S. Students.
- **NODET Position Scholarship for PhD Degree**, Sharif University of Technology.
- **Ranked 1st Among Undergraduate Students**, EE Dept., Shahid Chamran University (SCU).
- **NODET Position Scholarship for Master Degree**, SCU.
- **Award of Pioneer Student in Electrical Engineering**, SCU 2014, 2015.

Selected Publications (Google Scholar)

Ali Hadi Zadeh, Mostafa Mahmoud, Ameer Abdelhadi, and Andreas Moshovos. Mokey: Enabling narrow fixed-point inference for out-of-the-box floating-point transformer models. In *Proceedings of the 49th Annual IEEE/ACM International Symposium on Computer Architecture*, ISCA '22, 2022.

Andreas Moshovos, **Ali Hadi Zadeh**, Isak Edo Vivancos, and Omar Mohamed Awad. Quantization for neural network computation, March 24 2022. **US Patent App.** 17/130,690.

Omar Mohamed Awad, Mostafa Mahmoud, Isak Edo, **Ali Hadi Zadeh**, Ciaran Bannon, Anand Jayarajan, Gennady Pekhimenko, and Andreas Moshovos. Fpraker: A processing element for accelerating neural network training. In *Proceedings of the 54th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO '21, 2021.

Ali Hadi Zadeh, Isak Edo, Omar Mohamed Awad, and Andreas Moshovos. Gobo: Quantizing attention-based nlp models for low latency and energy efficient inference. In *Proceedings of the 53rd Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO '20, 2020.

Mostafa Mahmoud, Isak Edo, **Ali Hadi Zadeh**, Omar Mohamed Awad, Jorge Albericio, and Andreas Moshovos. Tensordash: Exploiting sparsity to accelerate deep neural network training. In *Proceedings of the 53rd Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO '20, 2020.

Ali Hadi Zadeh, Zissis Poulos, and Andreas Moshovos. Deep learning language modeling workloads: Where time goes on graphics processors. In *Proceedings of the Annual IEEE International Symposium on Workload Characterization*, IISWC '19, 2019.

Ali Hadi Zadeh, Matin Hashemi, Mohammad Labbaf, and Mostafa Parniani. A matrix-inversion technique for fpga-based real-time emt simulation of power converters. *IEEE Transactions on Industrial Electronics*, 66(2):1224–1234, 2019.

Teaching Experience

- **Course Instructor**
  University Of Toronto *2020-Present*
  - **APS1070:** Data Analytics and Machine Learning,
- **Teaching Assistant**
  University Of Toronto *2018–Present*
  - **Machine Learning**
    - **APS 360, MIE 1517:** Deep Learning, *Head TA*
    - **APS 1070:** Data Analytics and Machine Learning, *Head TA*
  - **Digital Systems**
    - **ECE 352, CSC 258, ECE243:** Computer Organization
    - **ECE 241:** Digital Systems, **ECE 334:** Digital Electronics, **CSC 367:** Parallel Programming

Selected Graduate Courses

- Machine Learning and Data Mining (A+)
- Parallel Programming and Architectures (A+)
- Reconfigurable Computing and FPGA Architecture (A+)
- Advanced Computer Architecture (A+)

Skills

- **ML:** NLP (Attention-based, Transformers, BERT-Family, LSTM), Computer Vision (CNN, ViT, GAN).
- **S/W & Tools:** Python/Pytorch, C/C++, MATLAB, CUDA, Git, Bash Scripts.