ALI HASSANI

ahassanijr@gmail.com | alihassanijr.com

alihassanijr | 🎓 Google Scholar | alihassanijr | alihassanijr

Atlanta, Georgia, United States.

Last update: July 28, 2025

Education

- **Georgia Institute of Technology** — *2024 - 2026 (expected)*
  PhD in Computer Science — Atlanta, GA.
  - Thesis: Reducing the $O(n^2)$ complexity of Attention at the Threadblock Level.
  - GPA: 4.00
- **University of Oregon** — *2021 - 2023*
  MS in Computer Science — Eugene, OR.
  - Thesis: Escaping the big data paradigm with compact transformers.
  - GPA: 4.22
- **University of Kerman** — *2016 - 2020*
  BS in Computer Science — Kerman, Iran.
  - Thesis: Clustering-based feature selection.
  - GPA: 3.81

Select Publications

[2025] Ali Hassani et al. **Generalized Neighborhood Attention: Multi-dimensional Sparse Attention at the Speed of Light** . *Preprint.*

[2024] Ali Hassani, Wen-Mei Hwu, and Humphrey Shi. **Faster Neighborhood Attention: Reducing the $O(n^2)$ Cost of Self Attention at the Threadblock Level** . In *Advances in Neural Information Processing Systems (NeurIPS).*

[2023] Ali Hassani et al. **Neighborhood Attention Transformer** . In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[2022] Ali Hassani and Humphrey Shi. **Dilated Neighborhood Attention Transformer** . *Preprint.*

[2021] Ali Hassani and Steven Walton et al. **Escaping the Big Data Paradigm with Compact Transformers** . *Preprint.*

Research Interests

- **Deep learning / AI architecture:** Developing high-performance and efficient neural network architectures, attention-based architectures.
- **High performance AI / ML systems:** Analytical performance models, performance optimizations (i.e. software pipelining, kernel fusion, and the like), developing new implementations / compute kernels.

Experience

- **SHI Labs at Georgia Tech** — *01/2024 - Present*
  Graduate Research Assistant — Atlanta, GA.
  - Conducted research in high-performance AI and ML systems.
  - Worked on improved software infrastructure for multi-dimensional sliding window / neighborhood attention: FNA (NeurIPS 2024) and GNA (under review, collaboration with NVIDIA).
  - Worked on a tensor parallelism solution within the NVIDIA CUTLASS framework (collaboration with NVIDIA).
- **NVIDIA Research** — *12/2024 - 07/2025*
  Research Intern — Remote position
  - Helped develop a parallelism strategy that scaled video / world generation to real-time level on a GB200 NVL72 rack, without any distillation, quantization, or sparsity.
  - Conducted research on sparse attention methods for accelerating Video / World Foundation Models on modern GPU architectures.
  - Developed **Generalized Neighborhood Attention (GNA)**, accompanied by an analytical performance model, and Fused Attention kernels for the Hopper and Blackwell architectures offering FLOP-proportional speedups.
  - Used a profiling approach to introduce GNA into the Cosmos Predict2 Video-to-World model, which results in up to **2.6X end-to-end inference speedup** with minimal loss in quality.

- **NVIDIA** *05/2024 - 08/2024*
  Software Performance Engineering Intern Remote position
  - Worked on low-latency matrix multiply (GEMM) kernels in CUTLASS for memory-bandwidth-bound LLM inference workloads.
  - Developed a Top-K and softmax GEMM fusion in CUTLASS targeting Mixture-of-Experts (MoE) workloads. Featured in NVIDIA developer blog on inline PTX as a performance optimization technique.
  - Worked on Distributed GEMM, a CUTLASS-native framework for running tensor parallel GEMMs. Featured in **GPU MODE** ▶.
- **HippoML** *06/2023 - 12/2023*
  Software Engineering Intern Remote position
  - Worked on bringing state-of-the-art Generative AI models to various hardware accelerators through system co-design.
  - Contributed to building the core engine, CUTLASS backend, and quantization solutions for attention and convolution.
- **SHI Labs at University of Oregon** *03/2021 - 12/2023*
  Graduate Research Assistant Eugene, OR.
  - Conducted research in computer vision and ML systems.
  - Developed Neighborhood Attention: a localized attention pattern bringing linear complexity and convolution-like behavior and inductive biases to attention.
  - Created and developed NATTEN: a PyTorch extension providing fast implementations of neighborhood and sliding window attention approaches.
  - Worked on Compact Transformers: mini vision transformers with state of the art image classification performance, trainable on limited data and compute budgets.
- **Picsart AI Research** *2022*
  Research Intern Remote position
  - Conducted research on training large-scale attention-based computer vision models.

Projects

- **NATTEN: Deep learning extension for multi-dimensional sliding window attention.** *2022 - Present*
  []
  - Offers fast kernels for local, dilated, causal, and strided forms of neighborhood attention, with an easy to use PyTorch interface.
  - Kernels cover all NVIDIA GPU architectures since Maxwell.
  - Ships fast arch-native kernels for the Hopper and Blackwell architectures, which can realize **theoretically maximum achievable** speedups (proportional to reduction in FLOPs over the fastest available kernels.)
  - Enables efficient training and inference for models built with neighborhood attention, and offers a variety of tools for different deep learning architectures, and performance analysis tools.
  - Applications range from classical computer vision tasks (image classification, object detection, image segmentation), to generative models (diffusion-based image, video, and world generation), prediction models (weather and climate forecasting), music structure analysis, and more.
  - Featured in **GPU MODE** ▶.
- **Neighborhood Attention Transformer: Efficient subquadratic vision transformers.** *2022 - 2023*
  []
  - Created hierarchical vision transformers that preserve translational equivariance, locality, and global inter-dependency modeling, with subquadratic attention complexity.
  - Pre-trained 20M to 200M parameter variants on image classification, extended to downstream tasks such as object detection and various image segmentation tasks.
  - Set a new state of the art score for some segmentation tasks at the time of publication.
  - Methodology extended later to image generation, and video/world generation.
- **Compact Transformers: Train vision transformers with limited data and compute.** *2021 - 2022*
  []
  - Provides pure PyTorch training recipes for very small vision transformers that can be trained on datasets as small as CIFAR-10, and even on consumer CPUs.
  - Achieved state of the art score on Flowers-102, and competitive scores on other datasets.
  - Preprint cited over 600 times.
  - Featured in Keras examples, and blog post featured in PyTorch's Medium.