Muhammad Haseeb

haseeb099m@gmail.com | github.com/ha405 | linkedin.com/in/muhammad-haseeb

RESEARCH INTERESTS

Efficient ML: Compression methods (pruning, quantization) to save compute and memory without loss in performance.

Domain Generalization: Adapting Models to distribution shifts through pruning to identify sparse subnetworks that preserve domain-invariant representations.

Mechanistic Interpretability: Analyzing internal circuit evolution to understand representational collapse in Federated Learning.

EDUCATION

Lahore University of Management Sciences (LUMS) Sep 2022 – May 2026
B.S. Computer Science
Relevant Coursework: Machine Learning, Deep Learning, Computer Vision, AI on Edge Devices, Advanced Topics in ML, LLM Systems, Linear Algebra, Probability.

PUBLICATIONS

– **BaCP: Backbone Contrastive Pruning for Preserving Representations in Extremely Sparse Neural Networks**.
Mohammad Haroon Khawaja, **Muhammad Haseeb**, Mohammad Fatim Shoaib, Muhammad Tahir. Submitted to *AAAI 2025*.

RESEARCH EXPERIENCE

Research Assistant Jan 2025 – Present
Centre for Urban Informatics, Technology and Policy (CITY), LUMS
Advisors: Dr. Muhammad Tahir, Dr. Zubair Khalid

– **Backbone Contrastive Pruning (BaCP):**
  ◦ Contributed to a generalized pruning framework to merge standard pruning criteria with contrastive learning to prevent representational collapse.
  ◦ Designed a **multi-objective loss function** that aligns sparse embeddings with pretrained, fine-tuned, and historical snapshot references to maintain feature consistency.
  ◦ Maintained accuracy comparable to dense baselines at up to **99% sparsity**, outperforming standard unstructured pruning approaches.
  ◦ Evaluated across CNNs, Vision Transformers, and Language Models on classification and masked language modeling tasks.

– **Domain Generalization:**
  ◦ Developed a framework to use visual queries for improving domain generalization performance in vision models.
  ◦ Trained visual queries via group relative query optimization and achieved a **3% accuracy** gain on PACS dataset compared to empirical risk minimization.
  ◦ Working on a pruning-based methodology to remove domain specific noise from network and obtain a stable and more generalized sparse subnetwork.

– **Quantization of Diffusion Models:**
  ◦ Training a MLP based on noise schedule to understand per timestep variance regardless of scheduler type and predict quantization parameters.
  ◦ Proposed a consistency based learning objective to align information flow across timesteps between quantized and unquantized models.

Research Assistant Jan 2025 – Aug 2025
Computer Vision & Graphics Lab, LUMS

- **KL Aware Quantization (KLAWQ):**
  - Proposed an augmented GPTQ framework that integrates knowledge distillation and supervised fine-tuning to improve low-bit quantization.
  - Combined the standard GPTQ Hessian with second-order curvature from KL-divergence and cross-entropy, better preserving teacher model outputs.
  - Achieved a **30% reduction in perplexity** compared to the standard GPTQ baseline at equivalent bit-widths by optimizing a combined MSE+KL+CE objective.
- **Single-Image Camera Calibration (SOFI-UGCL):**
  - Developed a hybrid method combining a Transformer front-end with geometric post-processing to recover the full camera projection matrix.
  - Used a Multi-Scale Deformable Transformer (SOFI) to predict geometric primitives (zenith point, horizon line) directly from image features.
  - Recovered intrinsic (K) and extrinsic (R, t) parameters by deriving the world origin from vanishing point intersections and enforcing rotation matrix orthonormality during training.

Research Intern May 2025 – Aug 2025
University of Illinois Urbana-Champaign (UIUC)

- Investigated LLM-based heuristics for automated `#ifdef` guard insertion in C codebases.
- Explored the use of large language models for software debloating and code dependency analysis.
- Contributed to the development of a VS Code extension supporting a C-to-Rust translation pipeline.

INDUSTRY EXPERIENCE

Machine Learning Engineer (Contract) Jul 2025 – Oct 2025
Innova Tech

- Automated a large-scale data annotation pipeline, reducing manual effort and improving dataset quality for object detection training.
- Improved model training pipelines and architectures, resulting in measurable gains in validation performance (approximately 4% accuracy improvement).
- Conducted ONNX and TensorRT-based quantization and inference optimization experiments, reducing memory footprint by up to approximately 40% while maintaining real-time inference on edge hardware.

TEACHING EXPERIENCE

Teaching Assistant Sep 2025 – Dec 2025
CS436: Computer Vision, LUMS

- Designed and supervised an Assignment on transfer learning using PyTorch, and on building a multithreaded object-detection pipeline in C++/OpenCV for real-time video streams.
- Designed and led 40 student groups in course project on building a virtual tour application using Structure from Motion (SfM) pipeline and state of the art computer vision techniques and tools.

SKILLS & OPEN-SOURCE

Technical Skills: Python, C, C++, Rust, SQL, TypeScript; PyTorch, Transformers, ONNX, TensorRT, diffusers, adapters.

Open-Source Contributions:

- **pytorch-image-models:** Added F1, precision, and recall metrics to evaluation and training pipelines for both single-GPU and distributed setups.
- **adapters:** Implemented PEFT support for Group Query Attention models, fixing tensor mismatch issues.