KARTIK SRINIVAS

kartiksr@cs.cmu.edu ◇ GH:kartiksrinivas007 ◇ LinkedIn ◇ +1(609)-781-8511

EDUCATION

Carnegie Mellon University - *MS, Machine Learning (Systems and Theory concentration)* — Pittsburgh, PA
GPA: **3.97**/4.00, Coursework: Deep RL, ML Systems, Database Systems, Advanced Stat Theory, PGM — *December 2025*

Indian Institute of Technology - *B.Tech Computer Science with a Minor in Artificial Intelligence* — Hyderabad, TN
GPA: **9.56**/10.00, Graduated with **Honors**, an **Institute Silver Medal** and **Triple Dean's List (2021-2024)** — *July 2024*

EXPERIENCE

Research Intern — Summer 2025
Waymo LLC — *Mountain View, CA*

- Developed **post-training** methods for Waymo **foundation models** that **improved collision-safety**
- Built **new model attention encoders** and launched training jobs for **billion-parameter autoregressive models**
- **Mined high quality datasets** from the driving logs with $< 0.1\%$ selectivity for improving open-loop collision avoidance
- Collaborated with **predictive planner** team to start improvment of **on-board Waymo planning models**
- **Awarded a peer-bonus** for writing efficient TensorProto parsers for a separate infrastructure team within Waymo

Machine Learning Research Intern — Summer 2023
UBC Vancouver & Vector Institute for AI — *Vancouver, BC*

- Pioneered **decentralized federated mutual learning**, enabling privacy-preserving, public-data-free learning
- **Improved state of the art** accuracies by 8 % by training 50 models across clients with heterogeneous data
- Invented a **distributed training algorithm** using latent alignment and deep differentially private image generation
- Collaborated with **Meta AI** to write a paper which was **accepted to ICML 2024**

PUBLICATIONS

Exact Unlearning via Model Merging at Scale — ICLR 2025
*K Kuo, A Setlur, **Kartik Srinivas**, A Raghunathan, V Smith*

Overcoming Data and Model Hetergeoneities in Decentralized Federated Learning — ICML 2024
*CY Huang, **Kartik Srinivas**, X Zhang, X Li*

Removing the Influence of Source Exclusive Classes in Domain Adaptation — Under Review
*A Devalapally, **Kartik Srinivas**, P Jain, VN Balasubramanian*

SKILLS

Frameworks: JAX, Tensorflow, Pytorch, Triton, Numpy, Scikit, Pandas, Wandb, Tensorboard, SLURM
Programming Languages: C++, Python, C, JavaScript, Julia, Matlab, RISC-V Assembly, x86-Assembly

PROJECTS

Triton Kernels for H-Nets and & Mamba - 2 (Mamba-2) — *Present*

- Note: This is present work at Dr. Albert Gu's Lab
- Developed GPU Kernels for the Exponential Moving average layer in H-nets **beating Mamba-2 GPU Kernel latency**
- Optimized memory accesses, block sizes and tiling to improve throughput of memory-bound sub-kernels
- Implemented alternative parallel prefix scan algorithms for chunked EMA recurrences in triton

DOSD: Decoupled Online Speculative Decoding (Poster) — *Spring 2025*

- Developed a novel speculative-decoding algorithm that **beats state of the art transformer inference latency**
- Asynchronously trained a draft-model copy on mistakes made during speculative decoding on another GPU
- Co-created a **novel weight serialization protocol** and utilized MPI for sending weights of draft-model across GPUs

NABLA: An ML Transpiler (Github) — *Fall 2022*

- Designed a new domain-specific programming language that **provides automatic differentiation primitives**
- **Invented a compile-time automatic differentiation engine** which traverses the syntax tree to backpropagate gradients
- Implemented parallelized tensor-algebraic rules for gradient calculation, **achieving a 30% speedup in backprop**
- **Led a 7-member-team** to deliver the full programming language white-paper, compiler & tests within 3 months

AWARDS & SERVICE

Awards: MITACS Research Scholar, Shastri-Indo-Canadian Institute Scholar, Google Research week awardee (150 students selected all over India), Institute Academic Excellence (Thrice), Institute Silver Medal, Microsoft Engage Hackathon Awardee
Teaching: Computability Theory, Convex Optimization, Deep Learning, Tensor Analysis & Introduction to Programming