Arnab Karmakar

+1 (206) 731-9687 • arnabk1@uw.edu • github.com/arnabk001 • linkedin.com/in/arnabk1 • website

Summary

- 4+ years of experience in developing end-to-end ML systems – from big data curation, algorithm development, model optimization, distributed training, benchmarking and deployment in production servers
- Research expertise in Multimodal AI, generative vision-language models and representation learning, with hands-on experience in LLMs, VLMs, Stable Diffusion, BERT and CLIP
- Skilled in optimizing GPU kernels (CUDA/HIP) for ML workloads and graphics shader pipelines (OpenGL)
- Experienced in leading teams and collaborating across orgs to deliver ML solutions for impactful applications

Skills

Tools/Tech	AWS, GCP, Docker, Python, C/C++, Git/Github, Linux, Shell Scripting, ROS, SQL, MATLAB, Tableau
Deep Learning	PyTorch, Tensorflow, Keras, DeepSpeed, Webdataset, Hugging Face, MLflow, RAY, OpenCV, scikit-learn, GPT, RLHF, RAG, Vector DB, Vision Transformers (ViT), BERT, CLIP
GPU Compute	CUDA, HIP, Nsight Compute, Nsight Graphics, Omniperf, ROCprof, Triton, TensorRT, OpenGL

Professional Experience

Machine Learning Research Intern, Brainchip Sep – Dec 2024

- Achieved a **3.5× improvement** in **model inference speed** and **8× reduction** in **embedding dimensions**, with no loss in accuracy, by implementing a novel representation learning algorithm into the State Space Model (TENNs)
- Improved performance across **generative language modelling (GPT-2)**, speech recognition and activity recognition benchmarks – validating cross-domain effectiveness of the representation learning method
- Enhanced model sparsity, interpretability, and downstream task performance for resource-constrained edge devices

Machine Learning Engineer (Intern), Truemedia.org Jun – Sept 2024

- Achieved **91%** accuracy in image deepfake detection and reduced false positives to only **3.5%** by developing a **reverse-search model**, integrating Google Cloud Vision, web scraping, and Large Vision-Language Models (LVLMs)
- **Deployed** the containerized deepfake detection model using **Docker**, implementing secure secrets management while optimizing latency for near real-time performance in production environments
- Developed an **explainable** image manipulation detector by using attention modules and multi-level feature fusion
- Implemented an end-to-end data processing pipeline – including data curation, de-duplication, and labeling – to catalogue the largest multimodal in-the-wild dataset of deepfakes

Senior Research Scientist, Indian Space Research Organisation Aug 2019 – Sept 2023

- **Led** a team of 6 to develop the **real-time on-board** astronaut **health anomaly detection** model, achieving 96.8% accuracy while minimizing false negatives to only 0.7%
- Developed a **time series analysis** model using **LSTM** to predict Remaining Useful Life (RUL) of aircraft turbojets from 26 sensor data streams, enhancing preventive maintenance operations
- **Coordinated** across 3 international organizations, 5 research centers and 15+ engineering teams to streamline 10+ space systems development, by managing technical requirements, platform constraints, and safety criticality
- Contributed in a 12-member team for Human-in-Loop usability evaluation, developed usability benchmarks and conducted A/B testing, providing data-driven insights resulting in improved human-system interface design
- Performed in-depth telemetry data analysis for 12+ space systems, pinpointing critical deviations and the root causes, providing design recommendations that significantly enhanced space mission safety

Research Experience

Graduate Research Assistant, Prof. Ranjay Krishna, RAIVN Lab (UW CSE) Sept 2024 – ongoing

- Implemented a large-scale distributed training method on GPU clusters for multimodal (image-text) datasets, scaling the training data from 3M to 200M data points
- Improved **chatbot response** accuracy by 30% (as per user surveys) using a **multi-agent** Retrieval-Augmented Generation (**RAG**) technique with Pinecone vector databases
- Developed a feature extraction pipeline from **Stable Video Diffusion (SVD)** model for temporal correspondence tasks, outperforming other self-supervised methods in motion understanding across multiple benchmarks

Research Experience

Graduate Research Assistant, Prof. Aravind KrishnaMurthy, SAMPL lab (UW CSE) Mar – June 2024

- Optimized the **Stream-K GEMM Kernel** on AMD MI100 GPUs, contributing to AMD's official open-source Composable Kernel Library
- Resolved performance bottlenecks by optimizing kernel launch configurations and implementing a 2-stage prefetching method, resulting in a 10% increase in L1 cache hit rate
- Optimized a standalone FlashAttention kernel using memory tiling, thread coarsening, and shared memory utilization to achieve a 3x speedup over naive GPU implementation while maintaining scalability upto 2048 tokens context length
- Utilized Nsight Compute/Systems for kernel profiling to identify and resolve performance bottlenecks – implemented advanced reduction techniques, mixed-precision arithmetic, and targeted kernel optimizations

Graduate Research Assistant, Prof. Deepak Mishra, Virtual Reality Lab (IIST) Jan – June 2019

- Designed a novel viewpoint invariant feature extraction and feature fusion model using Generative Adversarial Networks (GANs), achieving 9.6% improvement in rank-1 accuracy and 16% improvement in mAP [publication]
- Developed an unsupervised ML model using Variational Autoencoders for large astronomy datasets, conducted Exploratory Data Analysis and visualization, achieving an exceptional 81.4% IoU score [publication]

Education

University of Washington (UW), Seattle, USA Sep 2023 – June 2025
Master of Science (MS) in Electrical and Computer Engineering (Specialization: ML / GPU Computing)

Indian Institute of Space Science and Technology Aug 2015 – May 2019
Bachelor of Technology (B.Tech.) in Electronics and Communication Engineering

Projects

Analysis of Universal Attacks on Blackbox LLMs Jan – Mar 2024

- Developed advanced prompt injection attacks to identify vulnerabilities in popular LLMs, achieving 81% success rate
- Analyzed the effectiveness of LLM safety features and guardrails against various types of adversarial attacks

Optimizing ML models for improved performance on edge devices Jan – Mar 2024

- Implemented pruning and quantization techniques to reduce ML model size by 8X, achieving optimal compression while retaining performance, for low power edge device applications

Attribute conditioned face image generation using diffusion model Sep – Dec 2023

- Implemented a conditional diffusion model from scratch with learned attribute embedding and cosine noise scheduler
- Achieved smooth interpolation of face attributes and one-shot image editing with a 17.3%

Electrical Substation detection from hyperspectral satellite images Mar – June 2023

- Implemented a U-Net based deep learning model and extensive data augmentation technique for hyperspectral images
- Achieved pixel-level segmentation of electrical substations in satellite images, achieving an F1 score of 87.9%

Publications

[1] Arnab Karmakar and Deepak Mishra. "A robust pose transformational GAN for pose guided person image synthesis". In: *Computer Vision, Pattern Recognition, Image Processing, and Graphics: 7th National Conference, NCVPRIPG 2019, Hubballi, India, December 22–24, 2019, Revised Selected Papers 7*. Springer Singapore. 2020, pp. 89–99.

[2] Arnab Karmakar and Deepak Mishra. "Pose invariant person re-identification using robust pose-transformation gan". In: *arXiv preprint arXiv:2105.00930* (2021).

[3] Arnab Karmakar, Deepak Mishra, and Anandmayee Tej. "Stellar cluster detection using gmm with deep variational autoencoder". In: *2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. IEEE. 2018, pp. 122–126.

[4] Ankit Choudhary, Deepak Mishra, and Arnab Karmakar. "Domain adaptive egocentric person Re-identification". In: *Computer Vision and Image Processing: 5th International Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part III 5*. Springer Singapore. 2021, pp. 81–92.