Dhanesh Baalaji Srinivasan

ds7636@nyu.edu | +1 9292932179 | Linkedin | Github | New York City, NY (Open to relocate)

EDUCATION

New York University, New York City, NY; Master of Science in Computer Science; **GPA**: 3.75/4; **Honors*:*** Merit-based scholarship

TECHNICAL SKILLS

Programming Languages: Python, C, C++, C#, SQL, JavaScript. **Deep Learning frameworks:** PyTorch, JAX, TensorFlow, Keras.
AI Engineering: Retriever-Augmented Generation (RAG) systems, Fine-tuning Large Language Models (LLMs), LLM inference optimization.
LLMs: LangChain, Hugging Face, Lightning. **Vector search:** Opensearch, Elasticsearch. **GPU:** CUDA, Triton. **Big Data:** Hadoop, Spark.
Databases: Postgres, SQL Server, MongoDB. **Full Stack and Cloud:** AWS, Angular, Django, ASP .NET. **DevOps:** Docker, Kubernetes.

WORK EXPERIENCE

LOCOMeX, Inc., New York City, United States (Remote) Feb 2025 - May 2025
Software Engineer and MLOps Intern | Django, AWS Lambda, DynamoDB, RDS, Postgres, Python, ECR, Docker

- Engineered a low-latency search autocomplete feature using AWS Lambda and RDS, improving search responsiveness by 70%.
- Containerized an XGBoost-based sanctions risk prediction model as a serverless AWS Lambda function, enabling scalable, low-latency inference within the compliance pipeline.

New York University, New York City, United States Jan 2024 - May 2025
Graduate Research Assistant - Brooklyn Application,Architecture,Hardware Lab| DARPA Project | C, Python, Assembly, ARM NEON

- Optimized systolic array configurations (Input/Weight Stationary) for efficient hardware acceleration of a CNN using SCALE-Sim.
- Integrated a Last-level Cache into a Spectrum sensing Processor simulator and created sweeps to obtain the optimal cache size.
- Modeled and introduced variable Common Bus delays to assess signal detection throughput under various latency constraints.
- Developed Power Spectral Density and Match filter kernels using ARM v8.2 NEON for real-time spectrum sensing computations.

Graduate Course Assistant - High Performance Machine Learning | Pytorch, CUDA, C

- Evaluated students' ability to apply PyTorch, CUDA and C by grading programming assignments focused on ML engineering.
- Conducted office hours and provided guidance on advanced topics such as CUDA model, distributed training, and FlashAttention.

Psiog Digital Private Limited, Chennai, India Nov 2020 - May 2023
Software Engineer | Angular, ASP .NET Core, ASP .NET Framework, ASP .NET MVC, Javascript, C#, SQL.

- Devised a scalable Bidding system that incorporated an AI voice assistant which generated 4000+ user registrations within a month.
- Crafted RESTful APIs, designed SQL scripts, and responsive User Interfaces resulting in a 30% increase in user engagement.
- Fixed critical Extract, Transform, and Load (ETL) pipeline issues, saving $200k in potential losses from data downtime.
- Managed CI/CD pipelines across multiple products, reducing deployment times by 25% and hence improving release frequency.

PROJECTS

NAS-SegNet - A Novel efficient neural network for Medical Image Segmentation | NYU and IBM | PyTorch, IBM AnalogNAS, AIHWKIT

- Designed NAS-SegNet, a lightweight 800K-parameter segmentation model using IBM AnalogNAS, achieving 0.58 IOU (digital model) on MONAI's nuclei dataset, closely matching U-Net performance with over 90% fewer parameters.
- Adapted IBM's classification-focused supernetwork for segmentation by replacing pooling and dense layers with transpose convolution-based upsampling layers, enabling pixel-wise prediction.
- Performed hardware-aware training using IBM AIHWKit by simulating analog non-idealities such as device noise, and conductance drift; achieved 0.40 IOU, demonstrating the model's robustness for deployment on analog accelerators.

LlamaLearn - Retriever-Augmented Generation (RAG) flow in AWS for Large Language Models (LLMs) | Amazon Web Services, Python.

- Architected a scalable RAG system using DPR for dense retrieval and NeuralHermes-2.5 (Mistral-7B) for generation, deployed via AWS EKS and ECR with OpenSearch for vector search and DynamoDB for user-specific metadata to enable personalized answering.
- Engineered a modular information retrieval pipeline featuring document chunking, DPR-based vectorization, and OpenSearch k-NN search, enabling low-latency, semantically accurate real-time question answering.
- Improved answer quality and reduced hallucinations by injecting top-k retrieved chunks into the LLM for context-aware generation.

Fine-tuned Llama 3.1 8B for Math Question Answering | **Deep Learning** | Pytorch, Numpy, unsloth, huggingface

- Fine-tuned LLaMA 3.1 8B for math question answering using Rank-Stabilized LoRA and structured prompt engineering, achieving 82.04% test accuracy which is a 9.4% relative improvement over the 75% baseline.
- Leveraged 4-bit quantization and Unsloth's memory-optimized training stack to fine-tune LLaMA 3.1 8B on a single GPU (Colab's T4), reducing VRAM usage while maintaining factual accuracy in mathematical reasoning.

Subreddit Recommendations and Sentiment analysis on Reddit data | Pyspark, DistilBERT, TF-IDF

- Analyzed 3.8M Reddit posts using PySpark, TextBlob and DistilBERT to perform large-scale sentiment classification.
- Developed a content-based subreddit recommendation system using TF-IDF vectorization and cosine similarity to rank subreddits by the volume of posts exceeding a semantic similarity threshold with the user query.