Ibrahim Bathusha, Thameem Abbas

Email : tabbas97@gmail.com
Linkedin: https://www.linkedin.com/in/thameemabbas/
Mobile : +1-857-313-0768

Machine Learning Engineer, Distributed Systems, LLMs, Backend

Professional Experience

Machine Learning + Cloud Engineer Chennai, India
HuBe.ai, Multicoreware Inc. *Jun 2021 - Jun 2022*

- Conducted research on innovative architectures for deep recurrent multi-modal neural networks using mixed precision training and DDP training using SLURM as the manager.
- Engineered a highly scalable, multi-cloud capable and efficient SaaS platform delivering exceptional throughput and sub-100 ms system latency, empowering businesses to leverage near real-time applications and providing a competitive edge. Employed a variety of cloud-native ideas including containerization, microservices, load-balancing, WAFs, and, middlewares
- Achieved 60% reduction in maximum request latencies by implementation and data-driven optimization of dynamic batching
- Spearheaded the development of a redis-backed streaming-model back-end architecture, culminating in a 300% improvement in throughput, achieved through meticulous decoupling of modules and elimination of critical bottlenecks, significantly enhancing efficiency and performance
- Successfully orchestrated a 30% reduction in cloud expenditures through strategic optimization of instance sizes, enhanced demand-planning, and efficiency improvements in container image spin-up times resulting in improved response

Machine Learning Engineer - Computer Vision Chennai, India
PoseAnalytiq, Multicoreware Inc. *Jun 2019 - May 2021*

- Optimized computer vision models to match or surpass SOTA benchmarks while achieving a 4x reduction in memory and compute requirements pre-quantization via a range of compression techniques including pruning, and knowledge distillation
- Designed and trained model architectures optimized for various target hardware platforms(hardware-model co-design), ranging from GPUs to DSPs improving performance by upto 2x over base architectures
- Developed custom parsers and converters between multiple ML training and inference frameworks including Darknet, Tensorflow, PyTorch, ONNX, nGraph(Intel), and TensorRT enabling the use of novel model architectures.
- Leveraged parallel multi-stream inference to boost performance by 40% on TensorRT and optimized CPU performance by 20%
- Built the SDK to operate across Windows, and Linux, platforms running inference on CPU(x86, AVX), GPU(CUDA), and dedicated edge inference accelerators - ARM NN and Qualcomm SNPE. Employed Python/C++ interoperability to achieve high performance routines.
- Established automated build and test pipelines on Bitbucket employing shadow testing and A/B testing ensuring stable development and improving developer throughput

Education

Northeastern University Boston, MA
Master of Science in Computer Science *May 2024*
*Relevant Courses: **Natural Language Processing(NLP), Machine Learning, Algorithms**, Cloud Computing, **Scalable Distributed Systems***

PSG College of Technology - Anna University Coimbatore, India
Bachelor of Engineering - Electrical and Electronics Engineering *Apr 2019*
Relevant Courses: Digital Signal Processing, Linear Algebra, Probability and Statistics, Data Structures and Algorithms

Skills Summary

Languages and Databases: Python, C/C++, Java, Javascript, Node.js, HTML, CSS, SQL, C# | MySQL, MongoDB, Redis

Operating Systems and tools: Windows, Linux / Unix | Pytorch, TensorFlow, OpenCV, NumPy, ONNX Runtime, TensorRT, CUDA, OpenVINO, SNPE, NLTK, Pandas, Polars, Sklearn, Keras, TFlite, FastAPI, Git, JIRA, Jenkins, Gitlab, Docker, Kubernetes, Kafka, boto3, hdf5, Java Swing, GTest, PyTest, Matlab, Shell Scripting(Bash, Powershell), Terraform, React

Services: AWS (EC2, ECR, S3, Lambda, API Gateway, CloudWatch, IAM, Kinesis Video Streams, Fargate, ECS, EKS, Sagemaker, CloudFormation, Auto scaling, CloudTrail, Config, VPC Management, Eventbridge), GCP

Disciplines: Computer Vision, Natural Language Processing, Distributed Systems, Deep Learning, Backend Development, Cloud Computing, Infrastructure as Code(IaC)

Other Experience

Graduate Teaching Assistant Boston, MA, USA
Khoury College of Computer Sciences *Sep 2022 - Present*

- Building Scalable Distributed Systems - CS6650, Fundamentals of Cloud Computing - CS6620, Mobile Application Development - CS5520, Networks and Distributed Systems - CS3700
- Worked under the guidance of Professors Dr. Saripalli Prasad, Dr. Tony Mullen, Dr. Daniel Feinberg, Dr. Alden Jackson, and Dr. Christo Wilson

Projects

Safe-Spotter (Fall 2023): Created a geo-location driven app to present time-referenced crowd-sourced threat levels in the user's neighborhood and potential destination. Designed to be anomaly resilient in data processing to reject user bias and promote fairness

Multi-region Distributed Ticket Booking System (Fall 2022): Designed and built a prototype of a distributed ticket-booking system with a REST API endpoint. The system was designed to be resilient to multi-node failures implementing distributed transactions, data-replication management, and Java RMI. A simple store-front was also created.

Slapscape (Fall 2023): An app for urban sticker art enthusiasts to share, search, and interact with location-tagged posts featuring images, tags, and user profiles. Developed by using relational database and node.js at the backend and web technologies including next.js in the front-end