Suhas Jayaram Subramanya

Ph.D Student
Carnegie Mellon University
Advisor: Prof. Greg Ganger

June, 2021
suhas@cmu.edu | j.s.suhas@live.in
Webpage : suhasjs.github.io
Github : www.github.com/suhasjs

Education

Carnegie Mellon University — Pittsburgh, PA
Doctor of Philosophy (Ph.D) in Computer Science — *Aug '19 - Present*

Research Interests

Scheduling, Machine Learning systems, Approximate Nearest Neighbor Search

Publications

FreshDiskANN: A Fast and Accurate Graph-Based ANN Index for Streaming Similarity Search
Aditi Singh, **Suhas Jayaram Subramanya**, Ravishankar Krishaswamy, Harsha Vardhan Simhadri
Under Review, 2021.

Pollux: Co-adaptive Cluster Scheduling for Goodput-Optimized Deep Learning
Aurick Qiao, Sang Keun Choe, **Suhas Jayaram Subramanya**, Willie Neiswanger, Qirong Ho, Hao Zhang, Gregory R. Ganger, Eric P. Xing
15th USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2021.

PACEMAKER: Avoiding HeART attacks in storage clusters with disk-adaptive redundancy
Saurabh Kadekodi, Francisco Maturana, **Suhas Jayaram Subramanya**, Juncheng Yang, K. V. Rashmi, Gregory R. Ganger
14th USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2020.

DiskANN: Fast Accurate Billion-point Nearest Neighbor Search on a Single Node
Suhas Jayaram Subramanya, Devvrit, Rohan Kadekodi, Ravishankar Krishaswamy, Harsha Vardhan Simhadri
Neural Information Processing Systems (NeurIPS), Vancouver, 2019.

BLAS-on-flash: An Efficient Alternative for Large Scale ML Training and Inference?
Suhas Jayaram Subramanya, Harsha Simhadri, Srajan Garg, Anil Kag, Venkatesh Balasubramanian
16th USENIX Symposium on Networked Systems Design and Implementation (NSDI), Boston, 2019.

Exploration for Multi-task Reinforcement Learning with Deep Generative Models
Sai Praveen Bangaru, **Suhas Jayaram Subramanya**, Balaraman Ravindran
NeurIPS Deep Reinforcement Learning Workshop, Barcelona, 2016.

Work Experience

Project Singularity, Microsoft
Research Intern — *Jun - Aug '21,* ***Seattle***

Currently working on understanding and exploiting job characteristics for efficient scheduling of deep learning workloads at scale. Singularity is Microsoft's infrastructure for AI training and inference workloads.

Project Turing, Microsoft
Research Intern — *May - Aug '20,* ***Seattle***

Developed *FreshDiskANN*, an ANNS system capable of serving thousands of concurrent query, insert and delete operations on billion-point datasets with millisecond-scale latency on workstation-class machines with NVMe SSDs. FreshDiskANN is expected to power the next-generation of Bing Enterprise Search deployed to millions of organizations worldwide.

Microsoft Research India
Research Fellow *Jul '17 - Jul '19,* ***Bangalore***

Designed and developed cost-effective scalable machine learning systems that power production pipelines for topic-modeling, extreme multi-label learning, deep relevance model training, and approximate nearest neighbor serving. Developed DiskANN, an ANNS system that powers the Bing Web Search using NVMe SSDs, delivering hundreds of billions of k-ANNS queries at sub-5ms latencies at 10x lower cost compared to competing solutions.

Google India
Software Engineering Intern *May - Jul '16,* ***Bangalore***

Developed annotation metrics and production pipelines to understand efficacy of personalization re-rankers.

Research Projects

Heterogeneity-aware DNN Schedulers
Advisors: Prof. Greg Ganger, Dr. Aurick Qiao *Jan' 21 - Present,* ***CMU***

Recent advancements have enabled new class of DNN schedulers (Pollux) capable of adapting job parameters to maximize utility of allocated resources. Can we extend these schedulers to exploit heterogeneity in accelerators to improve cluster utility? Can we exploit transfer learning to generate performance profiles for incoming jobs?

Patents

Building a graph index and searching a corresponding dataset
Suhas Jayaram Subramanya, Ravishankar Krishaswamy, Harsha Vardhan Simhadri
US Patent App. 16/582,682, 2020.

Open Source Contributions

DiskANN
Microsoft Research, ≈18000 LOC [Github]

Open source implementation of DiskANN and Vamana algorithms in C++ for both Linux and Windows. DiskANN supports building and serving of SSD-based indices for k-ANNS queries on `uint8, int8,` and `float` datasets.

BLAS-on-flash
Microsoft Research, ≈10000 LOC [Github]

Open source implementation of BLAS-on-flash framework and runtime in C++ with sample kernels for matrix operations like `gemm` and `csrmm` and utility kernels like `sort` and `map-reduce`.

Importance Sampling for Learning Edge Topics (ISLE)
Microsoft Research, 2000+ LOC [Github]

Implemented a bounded-memory symmetric eigensolver using the Block Krylov-Schur algorithm for ISLE using the BLAS-on-flash framework that enables training of >10x larger models on workstation-class machines.

Teaching Experience

Introduction to Machine Learning
Instructor: Prof. Balaraman Ravindran *Jan - Apr '17,* ***IIT Madras***

Teaching Assistant on a MOOC hosted on NPTEL with over 6000 registered students. Course contents now archived to allow others to take the course at their own pace.