

Suhas Jayaram Subramanya

Ph.D Student

Carnegie Mellon University

Advisors: *Prof. Greg Ganger*

October, 2019

suhas@cmu.edu | j.s.suhas@live.in

Webpage : suhasjs.github.io

Github : www.github.com/suhasjs

EDUCATION

Carnegie Mellon University

Doctor of Philosophy (Ph.D) in Computer Science

Pittsburgh, PA

Aug '19 - Present

Indian Institute of Technology Madras

Bachelor of Technology (B.Tech) in Computer Science and Engineering

Chennai, India

Jul '13 - Jul '17

GPA: 9.32/10, Department Rank - 3 (Top 10%)

RESEARCH INTERESTS

Domain-specific systems (ML/Graph processing), Machine Learning

PUBLICATIONS

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.

BLAS-on-flash: an alternative for training large ML models?

Suhas Jayaram Subramanya, Srajan Garg, Harsha Simhadri
SysML Conference, Stanford, 2018.

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

Microsoft Research India

Research Fellow

Jul '17 - Jul '19, Bangalore

Worked on disk-based cost-effective, scalable machine learning systems. Resulting work deployed in production pipelines for topic-modeling, extreme multi-label learning, deep relevance model training, and approximate nearest neighbor serving.

Google India

Software Engineering Intern, Strategic Technologies team

May - Jul '16, Bangalore

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

Hyperverge Technologies Inc.

Algorithms Engineer

Jan - Dec '15, Chennai

Developed tools for album clustering and captioning using scene, timeline and geographic information obtained from a collection of photos and state-of-the-art scene labeling algorithms.

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.

RESEARCH PROJECTS

DiskANN - Disk-based Fast, Production-grade k-ANN Systems

Advisors: Dr. Harsha Simhadri, Dr. Ravishankar Krishnaswamy

Dec' 18 - Ongoing, Microsoft Research

Developed a novel disk-based system to serve k-approximate nearest neighbor queries on 1B points in ≈ 32 GB RAM. System to be deployed in production for Bing Web Search, improving quality of results for ≈ 150 billion queries a year. Currently working on adapting k-ANN indices to support insertions/deletions.

BLAS-on-flash - Disk-based Large Scale ML Training and Inference?

Advisor: Dr. Harsha Simhadri

Aug '17 - Jun '18, Microsoft Research

Conceptualized, designed and developed a framework to express matrix-based algorithms as dynamic computation graphs with nodes performing compute on a small subset of input data. Developed a high-performance runtime to execute these graphs on flash-resident data within a memory budget. Implementations of algorithms with a wide range of compute-communication ratios achieve performance parity with their in-memory variants at a substantially lower memory footprint, making 10x scalability in input sizes feasible.

Navigation for Muti-task Reinforcement Learning

Advisor: Prof. Balaraman Ravindran

Aug - Nov '16, IIT Madras

This project explores the problem of navigation in a distribution of maze-like environments with POMDP-like behaviours. We use a Variational Autoencoder in conjunction with a Gaussian Restricted Boltzmann Machine to model the agent's belief over the environment distribution and incentivise the agent to reduce uncertainty in the belief. Rollouts are used for exploration with Q-learning as the core learning algorithm.

Continuous Control for Simulated Creatures using Hierarchy of Policies

Advisor: Prof. Balaraman Ravindran

Jan - Jun '17, IIT Madras

Explored a hierarchical approach to continuous control inspired by the Encapsulation-Syllabus-Pandemonium (ESP). With influences from Feudal Reinforcement Learning, DDPG, and A3C, policies are organized in a hierarchy to learn increasingly abstract *sub-routines* through a programmer-defined curriculum.

Learning Input Conditional Language Models for Natural Language Generation

Advisor: Prof. Sutanu Chakraborti

Aug - Nov '16, IIT Madras

This work explores neural network architectures for learning surface realization, sentence planning, and content determination in an end-to-end manner from raw-data to full textual descriptions. Augmenting language-modeling LSTMs with an attention-layer over word-forms of inputs in the Prodigy-METEO dataset allows the model to *copy-paste* tokens from inputs in the final output.

OPEN SOURCE CONTRIBUTIONS

BLAS-on-flash

Microsoft Research, ≈ 10000 LOC

[\[Github\]](#)

Implemented the BLAS-on-flash framework in C++ with template-support. A high-performance multi-threaded runtime implements a custom caching layer and uses Linux kernel asynchronous I/O support for block I/O on SSDs with callbacks. Matrix-multiplication kernels like `gemm` (dense-dense) and `csrmm` (sparse-dense), and utility kernels like `csrscsc` (sparse-transpose), `sort` (Parallel Sample Sort), and `map-reduce` are implemented in the framework.

Importance Sampling for Learning Edge Topics (ISLE)

Microsoft Research, 2000+ LOC

[\[Github\]](#)

Implemented a symmetric eigensolver using the Block Krylov-Schur algorithm and ported memory-limited sections of ISLE to use the BLAS-on-flash framework. This new implementation is capable of training > 10 x larger models on multi-core workstation-class machines in the same memory envelope.