Supun Nakandala

Phone: (+1) 812-558-6888
Email: snakanda@eng.ucsd.edu
Web: scnakandala.github.io

3232 EBU3B CSE
9500 Gilman Drive
La Jolla, CA 92093

Research Interests My research interest lies broadly in the intersection of Systems and Machine Learning, an emerging area which is increasingly referred to as *Systems for ML*. In this space I operate as a data management researcher. Taking inspirations from classical data management techniques, I build new abstractions, algorithms, and systems to improve efficiency, scalability, and usability of machine learning workloads.

Education

University of California, San Diego, CA — Sept 2017 - Present
PhD, Computer Science. GPA: 3.97/4.00
Thesis Advisor: Prof. Arun Kumar
Courses: Database System Implementation, Advanced Data Analytics Systems, Data Models in Big Data Era, Advanced Compiler Design, Machine Learning, Recommender Systems and Web Mining, Algorithm Design and Analysis, Programming Languages

University of Moratuwa, Sri Lanka — Aug 2010 - April 2015
Bachelor of the Science of Engineering, Computer Science & Engineering.
GPA: 4.11/4.20.
Department Topper and Gold Medalist

Conference Publications

Cerebro: A Data System for Optimized Deep Learning Model Selection
Supun Nakandala, Yuhao Zhang, and Arun Kumar
VLDB 2020

A Tensor Compiler for Unified Machine Learning Prediction Serving
Supun Nakandala, Karla Saur, Gyeong-In Yu, Konstantinos Karanasos, Carlo Curino, Markus Weimer, and Matteo Interlandi
USENIX OSDI 2020

Vista: Declarative Feature Transfer from Deep CNNs at Scale
Supun Nakandala and Arun Kumar
ACM SIGMOD 2020

Extending Relational Query Processing with ML Inference
Konstantinos Karanasos, Matteo Interlandi, Doris Xin, Fotis Psallidas, Rathijit Sen, Kwanghyun Park, Ivan Popivanov, **Supun Nakandala**, Subru Krishnan, Markus Weimer, Yuan Yu, Raghu Ramakrishnan, Carlo Curino
CIDR 2020

Incremental and Approximate Inference for Faster Occlusion-based Deep CNN Explanations
Supun Nakandala, Arun Kumar, and Yannis Papakonstantinou
ACM SIGMOD 2019
Honorable Mention for Best Paper Award
Invited to ACM TODS 2020
Invited to ACM SIGMOD Research Highlight 2020

Gendered Conversation in a Social Game-Streaming Platform
Supun Nakandala, Giovani Cimpaglia, Norma Su, and Yong-Yeol Ahn
AAAI ICWSM 2017

Apache Airavata Security Manager: Authentication and Authorization Implementations for a Multi-Tenant eScience Framework
Supun Nakandala, Hasini Gunasinghe, Suresh Marru, and Marlon Pierce
IEEE e-Science 2016

Workshop and Demo Publications

Compiling Classical ML Pipelines into Tensor Computations for One-size-fits-all Prediction Serving
Supun Nakandala, Gyeong-In Yu, Matteo Interlandi, and Markus Weimer
NeurIPS 2019 MLSys Workshop

Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems
Supun Nakandala, Yuhao Zhang, and Arun Kumar
ACM SIGMOD 2019 DEEM Workshop

Demonstration of Krypton: Optimized CNN Inference for Occlusion-based Deep CNN Explanations
Allen Ordookhanians, Xin Li, **Supun Nakandala**, and Arun Kumar
VLDB 2019 Demo | SysML 2019 Demo

Materialization Trade-offs for Feature Transfer from Deep CNNs for Multimodal Data Analytics
Supun Nakandala, Arun Kumar
SysML 2018 Short paper

Anatomy of the SEAGrid Science Gateway
Supun Nakandala, Sudhakar Pamidigantam, Suresh Marru, Marlon Pierce
NSF XSEDE 2016

Pre-Prints

Cerebro: A Layered Data Platform for Scalable Deep Learning
Arun Kumar, **Supun Nakandala**, Yuhao Zhang, Side Li, Advitya Gemawat, and Kabir Nagrecha
Under Submission

Deep Learning Algorithms for Identifying Sedentary Behavior from Hip Worn Accelerometer Data
Supun Nakandala, Marta Jankowaska, Fatima Tuz-Zahra, John Bellettiere, Arun Kumar, and Loki Natarajan
Under Submission

Work Experience

Software Engineering Intern June 2020 - Sept 2020
AWS Redshift
Mentor: Yannis Papakonstantinou
Designed and implemented components of a confidential project.

Research Intern - Systems for ML June 2019 - Sept 2019
Microsoft Gray Systems Lab
Mentors: Matteo Interlandi, Markus Weimer
Designed and implemented Hummingbird system, a compiler for translating classical machine learning pipelines into tensor computations for unified and faster scoring of machine learning models.

Research Software Developer Oct 2015 - Aug 2017
Science Gateways Research Center - Indiana University
Manager: Marlon Pierce
Developed Apache Airavata, which is a software framework to compose, manage, execute, and monitor large scale applications and workflows on distributed computing resources such as local clusters, computational grids, and computing clouds.

Teaching Experience

Teaching Assistant - Systems for Scalable Analytics — UCSD - Winter 2020
Teaching Assistant - Advanced Data Analytics Systems — UCSD - Spring 2019

Scholarships and Awards

NSF travel award to attend ACM SIGMOD 2019 — NSF - 2019
Gold Medal for the Best Academic Performance — University of Moratuwa - 2015
Travel award to attend South Asia Workshop on Research Frontiers in Computing — National University of Singapore - 2014
Mahapola Higher Education Merit Scholarship — Govt. of Sri Lanka - 2010

Research Impact

Microsoft open-sourced Hummingbird system and uses it in ONNX ML Tools 2020
Ideas from project CEREBRO integrated into MADlib/Greenplum by VMWare 2019
CEREBRO system is being used by behavioral science researchers at UC San Diego medical school 2019
"Gendered Conversation in a Social Game-Streaming Platform" paper gains lot of media attention and creates awareness about the bleak issue of sexism in online game streaming platforms 2017
APACHE AIRAVATA science gateways middleware and the SEAGRID science gateway are widely used by computational science researchers to execute and manage computational jobs on university clusters and national supercomputing infrastructure 2017

Ongoing Projects

Project: Cerebro — Started September 2018

Deep neural networks are revolutionizing many ML applications. But there is a major bottleneck to wider adoption: the pain of *model selection*. This empirical process involves exploring the deep net architecture and hyperparameters, often requiring hundreds of trials. Alas, most ML systems focus on training one model at a time, reducing throughput and raising costs; some also sacrifice reproducibility. We are developing Cerebro, which is a system to raise deep net model selection throughput at scale and ensure reproducibility. CEREBRO uses a novel parallel execution strategy we call *model hopper parallelism* which is inspired by the multi-query optimization technique. Experiments on *Criteo* and *ImageNet* datasets show CEREBRO offers up to 10X speedups and improves resource efficiency significantly compared to existing systems like Parameter Server, Horovod, and task-parallel tools.

Project: Medical Data to Knowledge — Started September 2018

In this joint project between UCSD CS department and UCSD Medical School, I develop new deep learning-based techniques for predicting human activity (e.g., sitting, standing, and stepping) from accelerometer data. The data is collected from a large cohort of patients who wore accelerometer devices for seven days of free living. The goal is to develop accurate methods to predict human activity from these accelerometer data and then use them in downstream human activity and metabolic health correlation analysis. The challenges of this project include working with large volumes of training data (1 TB) and performing extensive *model selection* such as neural architecture search and hyperparameter tuning.

Technical Talks

Cerebro: A Data System for Optimized Deep Learning Model Selection
VLDB 2020; SIGMOD 2019; UCSD CNS Research Review 2019

Vista: Optimized System for Declarative Feature Transfer from Deep CNNs at Scale
SIGMOD 2020; UCSD CNS Research Review 2018

Incremental and Approximate Inference for Faster Occlusion-based Deep CNN Explanations
ACM SIGMOD 2019

A Tensor Compiler for Unified Machine Learning Prediction Serving
Microsoft Gray Systems Lab 2019