Anthony Hitchcock Thomas

530-781-3801

3241 Via Marin, La Jolla, CA 92037

ahthomas@eng.ucsd.edu ~ https://github.com/thomas9t

Research and Publications

Hierarchy-Aware Machine Learning Inference for Networked Sensing Applications - **Anthony Thomas**, Yunhui Guo, Yeseong Kim, Baris Aksanli, Arun Kumar, and Tajana Rosing

- To Appear: ICNSC 2019
- Data communication is a significant bottleneck for many IoT based Machine Learning (ML) applications. Prior work has proposed a "hierarchical" approach to ML in IoT wherein edge devices compute successively more abstracted (e.g. compressed) representations of input data. While this reduces communication it complicates ML deployment as the cloud can no longer access raw feature values. We show that several popular models are infeasible in this setting and propose communication-efficient alternatives. We motivate our proposed models using ML theory and present detailed evaluations on three real-world sensor based ML tasks. We show our approach leads to up to 67% decrease in energy use and 63% reduction in latency while preserving accuracy.

A Comparative Evaluation of Systems for Scalable Linear Algebra-Based Analytics - **Anthony Thomas** and Arun Kumar

- To Appear: VLDB 2019
- Relational algebra has proved an exceptionally versatile and powerful tool for decades of research into relational data management and query optimization. The past several years have seen growing interest in using linear algebra to similar purpose in machine learning and other forms of computationally demanding matrix based mathematics. This novel form of linear-algebra based program optimization, combined with a resurgence of interest in distributed numerical computing using tools like Apache Spark and Google's TensorFlow, has given rise to several new tools which attempt to apply the lessons learned from relational data management to distributed linear algebra based programs. In this work we experimentally compare several such systems and provide insight into their relative strengths and weaknesses.

Cache Optimized Model Serving with Resource Constraints - **Anthony Thomas**, Niketan Pansare and Berthold Reinwald

- Under preparation for VLDB 2019
- Modern enterprise relies heavily on sophisticated machine learning (ML) models to generate specialized content for users. Performing inference at production scale while meeting service objectives on latency presents significant systems challenges. We present a system which casts model serving as a multi-producer multi-consumer scheduling problem and appeal to results from queuing theory to design a scheduler which maximizes throughput subject so a soft constraint on latency in a resource constrained environment. Our scheduler builds on prior work from TCP congestion avoidance and task scheduling for map-reduce to design an on-line algorithm which determines optimal batch size and compute device given the cache locality of model weights.

Grants

Measuring Local Area Income Using Neural Networks Trained on Satellite Imagery - Arman Khachiyan, **Anthony Thomas**, Luke Sanford, Jennifer Burney, Alexander Clonninger, Gordon Hanson, and Tajana Rosing

- $\$175,000$ - Supported by the Russel Sage Foundation

Patents

A Storage Architecture for Heterogeneous Multimedia Data - with Dangyi Liu, Kai Zhou, Yidi Zhang, and Ruiliang Zhang

- Patent Pending
- We describe a system for distributed storage and querying of heterogeneous multimedia data gathered from autonomous driving vehicles.

An HTTP Based Caching and Authentication Scheme - with Dangyi Liu, Kai Zhou, Yidi Zhang, and Ruiliang Zhang

- Patent Pending
- We describe a system for accessing data stored in a distributed environment which combines a robust logging and authentication procedure with a fine grained disk based cache management system.

Education

The University of California, San Diego
PhD - Computer Science and Engineering
System Energy and Efficiency Lab - PI Tajana Rosing
3.9/4.0 GPA

- **Relevant Coursework**: Computer Vision I-III (Multiple View Geometry), Analysis of Algorithms, Convex Optimization, Universal Probability and Applications, Statistical Learning, Information Theory, Principles of Database Systems, Database System Implementation, Topics in Advanced Analytics, Advanced Compiler Design, Robotic Software.

Brown University

- **Relevant Coursework**: Machine Learning, Mathematical Statistics
- All coursework taken alongside full time employment.

The University of California, Berkeley
BS with High Distinction - Agricultural and Resource Economics (2013) 3.91/4.00 GPA

- **Relevant Coursework**: Statistical Computing, Linear Algebra and Differential Equations, Calculus I-III, Probability and Statistics, Econometrics I-II, numerous economic theory courses.
- **Academic honors**: Phi Beta Kappa, High Distinction in the College, High Honors in the University.
- **Other Activities**: California Lightweight Crew - Team Captain and President (2013).

Employment

IBM Research - Almaden *Summer Research Intern* $\sim$ June 2018 - September 2018

- Worked on design and implementation of a novel system for machine learning model serving. Developed a scheduling algorithm to optimize model serving in resource constrained environment.
- Contributed to development of GPU execution framework for Apache SystemML - an open source platform for scalable machine learning.
- Work is currently under preparation for publication

TuSimple LLC *Systems and Infrastructure Engineer Intern* $\sim$ June 2017 - September 2017

- Worked with a team of engineers to design and implement a customized distributed storage and querying system for sensor data streams produced by self driving cars.
- Developed a system to improve data access efficiency by load balancing across a pool of cache servers. Implemented a custom HTTP based data access and authentication/logging scheme with a transparent cache management system.
- Developed a tool to robustly transfer large data files over unstable network connections.

Brown University
Research Analyst: Professors Nathaniel Baum-Snow and Justine Hastings $\sim$ July 2013 - July 2015
Senior Research Analyst: Professors Justine Hastings and Jesse Shapiro $\sim$ July 2015 - July 2016

- Worked with Brown faculty members to conduct research in theoretical and applied economics for academic publication. Primarily responsible for the design and implementation of code to store, analyze and visualize research data.
- Designed and implemented a database to store approximately 1.5TB of data from a large nationwide retailer for use in quantitative research projects. The database implementation significantly reduced run-time compared to a previous file based implementation using Python and HDF5.
- Improved the efficiency of analysis code by implementing statistical models in Numpy using low-level linear algebra routines. The Python implementation enabled estimation of models that were computationally infeasible using Stata.

US Department of Agriculture - Forest Service
Wilderness Ranger Intern $\sim$ May 2011 - September 2011

- Worked independently and in small crews doing backcountry trail maintenance and surveying. Tours lasted 5-10 days in wilderness and all maintenance was performed using only hand operated tools.