# Yuhao Zhang

CONTACT INFORMATION

EBU3B 3230
CSE Department
UC San Diego
La Jolla, CA 92093 USA

*E-mail:* yuz870@eng.ucsd.edu
*Web:* https://yhzhang.info/

EDUCATION

**University of California, San Diego**, La Jolla, California USA

Ph.D. Student, Computer Science, September 2017 - December 2022 (anticipated)
- Dissertation Topic: "High-throughput Data Systems for Deep Learning Workloads"
- Advisor: Prof. Arun Kumar

**Nankai University**, Tianjin, China

B.S., Theoretical Physics, June 2016
- Advisor: Prof. Xueqian Li and Prof. Yangang Miao

RESEARCH INTERESTS

My research is primarily on machine learning systems, including systems powered by applied ML that enable novel applications and systems designed for ML to make data science easier and faster. Past and ongoing projects include systems for: distributed deep learning model selection and training, distributed in-database deep learning, distributed large-scale graph neural network training, and video analytics and querying.

PROJECTS

**Distributed High-Throughput Graph Neural Network Training System (Lotan):**

Currently work in progress, Lotan is a system designed for high-throughput, large-scale graph neural network (GNN) training. It aims to solve the scalability bottlenecks many current systems face through a novel fusion of graph processing systems and deep learning systems. It proposes many novel techniques, including a cost-based query optimizer, a graph partitioning scheme, and a share-memory-based IPC architecture to make such fusion possible and efficient.

**Distributed Deep Learning on Data Systems (Cerebro-DS):**

Deep learning (DL) is growing in popularity for many data analytics applications, including among enterprises. Large business-critical datasets in such settings typically reside in RDBMSs or other data systems. The DB community has long aimed to bring machine learning (ML) to DBMS-resident data. Given past lessons from in-DBMS ML and recent advances in scalable DL systems, DBMS and cloud vendors are increasingly interested in adding more DL support for DB-resident data. In this paper, we show that there is no single "best" approach to achieve that goal and an interesting tradeoff space of approaches exists. We explain four canonical approaches, compare them analytically on multiple criteria (e.g., runtime efficiency and ease of governance) and compare them empirically with large-scale DL workloads. Our experiments and analyses show that it is non-trivial to meet all practical desiderata well and there is a Pareto frontier; for instance, some approaches are 3x-6x faster but fare worse on governance and portability. Our results and insights can help DBMS and cloud vendors design better DL support for DB users.
*Project homepage:* `https://adalabucsd.github.io/cerebro.html`
*Open-sourced release:* `https://github.com/makemebitter/cerebro-ds`

**Resource-efficient Distributed Deep Learning Model Selection and Training System (Cerebro):**

Cerebro is a layered data platform for scalable deep learning. One of the biggest challenges for scalable deep learning is model selection, which is difficult and resource-heavy. This system provides high-level APIs for deep learning model selection and optimizes distributed deep learning training for model selection workloads. It adopts a new form of parallelism called Model Hopper Parallelism (MOP) and is the resource-optimal choice. It can expedite distributed deep learning training by 3x-10x compared to Horovod and TensorFlow Parameter Server. There are also subsequent works about bridging data systems (such as Apache Spark, Distributed databases, etc.) with distributed DL systems.
*Project homepage:* `https://adalabucsd.github.io/cerebro.html`
*Open-sourced release:* `https://adalabucsd.github.io/cerebro-system`

**Video Querying System under Unbounded Vocabulary Setting (Panorama):**

Panorama is a video querying system for object detection and classification. It is designed to tackle the life-long learning or bounded vocabulary issues. Computer-vision-based video analytics systems often require *update* of the vocabulary as new classes/labels need to be added from time-to-time. Such updates involve re-training of the deep learning model, which requires expertise and is very time-consuming. Panorama is built to partially avoid and automate this process while expediting inference speed at the same time.
*Project homepage:* `https://adalabucsd.github.io/panorama.html`
*Open-sourced release:* `https://github.com/makemebitter/Panorama-UCSD`

| RESEARCH IMPACT | | |
|---|---|---|
| Cerebro and MOP integrated into Greenplum Database and shipped by VMware | | 2019 |
| Code of Cerebro integrated into Apache MADlib project | | 2019 |

ACADEMIC EXPERIENCE

**University of California, San Diego**, La Jolla, California USA

*PhD Student* **Sept 2017 - present**
Includes current Ph.D. research, Ph.D. and Masters level coursework and research/consulting projects.
Courses taken: Machine Learning, Deep Learning, Data Mining & Analytics, Advanced Data Analytics, Computer Vision, Database Systems, Advanced Algorithms, Advanced Compilers, Principles of Programming Languages, Introduction to Robotics

*Teaching Assistant for CSE234: Data Systems for Machine Learning* **Winter 2021**
Helped designing and supervising a research-oriented course on machine learning systems. Mentored 12 master students with their course projects ranging from advanced implementation of cutting-edge research, to evaluation and surveying the state-of-art work, and to open-ended research.

*Teaching Assistant for DSC102: Systems for Scalable Analytics* **Winter 2020**
Developed the first edition of course assignments and auto-grading programs with Python and Bash. The assignments involve Python Dask, Spark, AWS EC2/S3/EBS, and Kubernetes. These assignments have been adopted by the course ever since and used by 500+ students.

**Texas A&M University**, College Station, Texas USA

*Research Intern* **Summer 2015**
Worked on vision-based object tracking, modeling, and data analytics.

**Institute of Physics, Chinese Academy of Sciences**, Beijing, China

*Research Intern* **Summer 2014**
Theoretical physics research. Particle physics.

**Nankai University**, Tianjin, China

*Research Assistant*                                                                                                    **2013 - 2016**
Theoretical physics research. Black holes, gravity, and neutrinos.

PROFESSIONAL        **Microsoft Gray System Lab**, California USA
EXPERIENCE
                   *Research Intern*                                                                                  **Summer 2021**
                   Worked on machine learning system research, focusing factorized in-DBMS machine learning.

                   **VMware**, Palo Alto, California USA

                   *Software Engineer Intern*                                                                         **Summer 2019**
                   Worked on the first in-DBMS deep learning system, allowing training and inference of deep learning
                   models with TensorFlow on database-resident data. Integrated my research project, Cerebro, into
                   the deep learning training infrastructure of Greenplum Database, boosting efficiency by over 10x.
                   Contributed to the Apache MADlib project in Python and SQL. Lead the development of a major
                   release. This project has been incorporated into Greenplum and production-ready for VMware's
                   customers.

                   **Opera Solutions**, San Diego, California USA

                   *Data Scientist Intern*                                                                            **Summer 2018**
                   Worked on a theatre scheduling & recommender system. Proposed new models and optimized the
                   existing system.

PUBLICATIONS       *Some Damaging Delusions of Deep Learning Practice (and How to Avoid Them)*
                   Arun Kumar, Supun Nakandala, Yuhao Zhang
                   KDD 2021 Deep Learning Day

                   *Distributed Deep Learning on Data Systems: A Comparative Analysis of Approaches*
                   Yuhao Zhang, Arun Kumar, Frank McQuillan, Nandish Jayaram, Nikhil Kak, Ekta Khanna, Orhan
                   Kislal, and Domino Valdano
                   VLDB 2021

                   *Cerebro: A Layered Data Platform for Scalable Deep Learning*
                   Arun Kumar, Supun Nakandala, Yuhao Zhang, Side Li, Advitya Gemawat, and Kabir Nagrecha
                   CIDR 2021

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

                   *Panorama: A Data System for Unbounded Vocabulary Querying over Video*
                   Yuhao Zhang and Arun Kumar
                   VLDB 2020

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

                   *Three-generation neutrino oscillations in curved spacetime*
                   Yuhao Zhang and Xueqian Li
                   Nuclear Physics B, 2016

*Self-regular black holes quantized by means of an analogue to hydrogen atoms*
Chang Liu, Yangang Miao, Yumei Wu, and Yuhao Zhang
Advances in High Energy Physics, 2016

| | | |
|---|---|---|
| PRESENTATIONS | *Distributed Deep Learning on Data Systems* | |
| | VLDB 2021 | **August 2021** |

| | | |
|---|---|---|
| | *A Layered Data Platform for Scalable Deep Learning* | |
| | UCSD Database Seminar | **October 2020** |
| | UCSD CNS Research Review | **October 2020** |
| | *Resource-Efficient Deep Learning Model Selection on Apache Spark* | |
| | Spark+AI Summit 2020 | **June 2020** |
| | UCSD Database Seminar | **May 2020** |
| | *Cerebro: A Data System for Optimized Deep Learning Model Selection* | |
| | VLDB 2020 | **September 2020** |
| | UCSD Database Seminar | **June 2020** |
| | *Panorama: Unbounded Vocabulary Queries on Video* | |
| | VLDB 2020 | **September 2020** |
| | UCSD Database Seminar | **June 2020** |
| | Poster presentation, UCSD CSE Research Open House | **January 2019** |
| | UCSD Database Seminar | **November 2019** |
| | Poster presentation, SoCal DB Day 2018 | **October 2018** |
| | *Apache MADlib 1.17 Showcase* | |
| | VMware and online | **September 2019** |

| | |
|---|---|
| SERVICE | **External reviewer**: SIGMOD 2020, SIGMOD 2021, VLDB 2022 |
| | **Reviewer**: JMLR MLOSS 2022 |

| | | |
|---|---|---|
| MISC | SIGMOD 2021 *Students and Postdocs in DB Panel Discussion* | **June 2021** |

| | |
|---|---|
| HONORS AND AWARDS | Best Thesis Award, School of Physics, Nankai University, 2016 |
| | Wang Kechang Scholarship for Academic Distinction, 2014 |
| | Gong-Neng Scholarship, 2013 |
| | Poling Academy Scholarship, 2012-2016 |

| | |
|---|---|
| TECHNICAL SKILLS | • Programming languages: Python, C/C++, Scala, Java, SQL, R |
| | • Data platforms: Spark, Dask, PostgreSQL, Greenplum Database, Hadoop, Hive |
| | • Machine learning frameworks: TensorFlow, PyTorch, Keras, xgBoost, LightGBM, Scikit-learn |
| | • Other data analytics packages: Pandas, NumPy, Matplotlib, OpenCV, Scikit-image, Pillow/PIL |
| | • Cloud/Cluster tools: AWS EC2/S3/EMR, Google Cloud, Azure, Kubernetes, Docker |
| | • Miscs: MPI, MATLAB, Mathematica, CERN ROOT, CERN GEANT4 |