Dong He

▪ dongheuw.github.io ▪ donghe@cs.washington.edu

EDUCATION

University of Washington, **Ph.D. in Computer Science** — USA
- Advisor: Prof. Magdalena Balazinska — Sep 2019 – present

Fudan University, **B.Sc. in Computer Science (Honors)** — China
- Cumulative GPA: 3.6 / 4.0, School Rank: 6 / 118 — Sep 2015 – Jul 2019

University of Birmingham, **Exchange Undergraduate Student** — UK
- First Class Honors' Grades — Sep 2017 – Dec 2017

PUBLICATIONS

Query Processing on Tensor Computation Runtimes [Paper] [Press I] [Press II] [Talk]
- ***D. He***, *S. Nakandala, D. Banda, R. Sen, K. Saur, K. Park, C. Curino, J. Camacho-Rodríguez, K. Karanasos, M. Interlandi.* VLDB 2022.

Share the Tensor Tea: How Databases can Leverage the Machine Learning Ecosystem [Paper]
- *Y. Asada*, V. Fu*, A. Gandhi*, A. Gemawat*, L. Zhang*,* ***D. He****, V. Gupta, E. Nosakhare, D. Banda, R. Sen, M. Interlandi.* VLDB 2022 Best Demo Award.

DeepEverest: Accelerating Declarative Top-K Queries for Deep Neural Network Interpretation [Paper] [Extended Technical Report] [Website] [Code] [Talk]
- ***D. He****, M. Daum, W. Cai. M. Balazinska.* VLDB 2022.

VOCAL: Video Organization and Interactive Compositional AnaLytics [Paper] [Website] [Talk]
- *M. Daum*, E. Zhang*,* ***D. He****, M. Balazinska, B. Haynes, R. Krishna, A. Craig, A. Wirsing.* CIDR 2022.

VSS: A Storage System for Video Analytics [Paper] [Technical Report] [Website] [Code]
- *B. Haynes, M. Daum,* ***D. He****, A. Mazumdar, M. Balazinska, A. Cheung, L. Ceze.* SIGMOD 2021.

TASM: A Tile-Based Storage Manager for Video Analytics [Paper] [Website] [Code]
- *M. Daum, B. Haynes,* ***D. He****, A. Mazumdar, M. Balazinska.* ICDE 2021.

Accelerating Mobile Applications at the Network Edge with Software-Programmable FPGAs [Paper]
- *S. Jiang,* ***D. He****, C. Yang, C. Xu, G. Luo, Y. Chen, Y. Liu, J. Jiang.* INFOCOM 2018.

Incorporating Location Based Social Networks in the Prediction of Real-Time Taxi Demand with Deep Learning [PDF]
- ***D. He****, Y. Chen.* Poster Session, CoNEXT 2018.

INVITED TALKS AND PRESENTATIONS

Query Processing on Tensor Computation Runtimes
- VLDB 2022, together with Matteo Interlandi [Video] — Sep 2022
- RelationalAI Seminar — Jun 2022

Share the Tensor Tea: How Databases can Leverage the Machine Learning Ecosystem
- VLDB 2022 Demonstration, together with Matteo Interlandi — Sep 2022

DeepEverest: Accelerating Declarative Top-K Queries for Deep Neural Network Interpretation
- VLDB 2022 [Video] — Sep 2022

SELECTED AWARDS

- Paul G. Allen Fellowship, UW CSE — 2019 – 2020
- Outstanding Undergraduate Graduates, Shanghai Region — 2019
- Honors Student Award, Top Talent Undergraduate Program, Fudan University — 2019
- Wangdao Scholar, Undergraduate Research Opportunities Program, Fudan University — 2018
- First Class Scholarship, Fudan University — 2016 – 2017
- First Prize, National Mathematical Contest in Modeling, Shanghai Division — 2016
- Silver Medal, ACM International Collegiate Programming Contest, Asia Regional — 2015

- Silver Medal, National Olympiad in Informatics, National Finals 2014
- First Prize, National Olympiad in Informatics in Provinces, Guangdong Division 2009 – 2014

TEACHING AND SERVICE

- Teaching Assistant, UW CSEP 590A: Machine Learning for Big Data Spring 2022
- Head Teaching Assistant, UW CSED 516: Scalable Data Systems and Algorithms Fall 2021
- Student Volunteer, VLDB 2020 Sep 2020

RESEARCH EXPERIENCE

Query Processing on Tensor Computation Runtimes Microsoft, UW

With Microsoft Gray Systems Lab Jun 2021 – Jun 2022

- Designed and implemented Tensor Query Processor (TQP), the first query processor that runs atop tensor computation runtimes (TCRs). TQP, consisting of a collection of novel tensor-based implementations for relational operators and a compiler stack, transforms SQL queries into tensor programs and executes them on TCRs.
- TQP supports the full TPC-H Benchmark. With TQP, we demonstrate that the tensor interface of TCRs is expressive enough to support all common relational operators. Meanwhile, TQP can support various hardware while only requiring a fraction of the usual development effort.
- Experiments show that TQP can improve query execution time by up to 10x over specialized CPU- and GPU-only systems. When machine learning model inference and SQL queries are used in concert, TQP is able to provide end-to-end acceleration for a 9x speedup over CPU baselines.

Accelerating Declarative Top-K Queries for Deep Neural Network Interpretation [Website] UW

Advisor: Prof. Magdalena Balazinska Oct 2019 – Apr 2021

- Designed, implemented, and evaluated DeepEverest, a system for the efficient execution of commonly-used *interpretation by example* queries over the activation values of a deep neural network.
- DeepEverest consists of an efficient indexing technique and an instance-optimal query execution algorithm, as well as several important optimizations.
- Experiments with our prototype implementation show that DeepEverest, using less than 20% of the storage of full materialization, significantly accelerates individual queries by up to 63x and consistently outperforms other methods on multi-query workloads that simulate DNN interpretation processes.

The VisualWorld Video Data Management Project [Website] UW

With the VisualWorld team Oct 2019 – present

- **VOCAL**: a vision of a video data management system that supports efficient data cleaning, exploration and organization for large-scale video data, as well as processing complex compositional queries, even when no pretrained model exists to extract semantic content.
- **TASM**: a tile-based storage manager for video data which enables spatial random access into encoded videos. TASM speeds up content retrieval queries by an average of over 50% and up to 94%, and also improves the throughput of the full scan phase of object detection queries by up to 2x.
- **VFS**: a system that decouples application design from video data's physical layout and compression optimizations. This decoupling allows application and system developers to focus on their relevant functionality, while VFS handles the low-level details associated with video data persistence. VFS also improves read performance by up to 54%, and reduces storage costs by up to 45%.

FPGA-Based Edge Computing for the Acceleration of Mobile Applications Peking University

Advisor: Prof. Chenren Xu Jul 2017 – Aug 2017

- Designed an FPGA-based edge computing model, which can effectively reduce the response time and energy consumption of interactive mobile applications.
- Implemented a proof-of-concept prototype, and conducted experiments in a case study using 3 computer vision-based interactive applications designed by us.
- Experimental results showed that our system can reduce the response time and execution time by up to $3\times/15\times$ respectively over CPU-based edge/cloud offloading and achieve up to $29.5\%/16.2\%$ improvement on energy efficiency on mobile device/edge nodes, respectively.

Improving the Prediction of Real-Time Taxi Demand with External Information Fudan University

Advisor: Prof. Yang Chen Sep 2018 – Jan 2019

- Proposed a deep learning-based approach which incorporates user check-in data from a Location-Based Social Network to improve the prediction of the taxi demand in different regions at different times.
- Integrated the taxi trip data with around 1 million user check-ins collected from the Swarm App. Evaluation on a dataset containing 35 million taxi trip records showed that our method achieves 21.27% lower MAPE and 6.96% lower RMSE compared to existing approaches.

INDUSTRY EXPERIENCE

Microsoft, **Research Intern** Remote

- With Gray Systems Lab. Jun 2021 – Sep 2021
- Worked on running relational queries on tensor computation runtimes.

Goldman Sachs, **Technology Summer Analyst** Hong Kong

- With the Product Accounting and Risk Analysis team. Jul 2018 – Sep 2018
- Global Winner for Goldman Sachs 2018 Intern Engineering Challenge.
- Re-designed and re-implemented the logic of the true-up job which reconciles the estimated profit and loss (PnL) with the actual PnL. My enhancements, deployed in production, considerably reduce the memory usage of the true-up job, which significantly reduces the chances of job failure.

Tencent, **Engineering Intern** Shenzhen

- With YouTu Lab led by Prof. Jiaya Jia and Prof. Yu-Wing Tai. Jan 2018 – Feb 2018
- Analyzed the liveness and dependencies of the nodes in neural networks, and reduced the memory consumption of such models in real-world products by memory sharing.
- Developed tools for the collection and annotation of large-scale image data, and collected massive data for training image classification models in real-world products.