Dong He

donghe@cs.washington.edu | dongheuw.github.io | linkedin.com/in/dongheuw | +1 (206) 295-6340

Education

University of Washington Sep 2019 – Jun 2024 (expected)
PhD in Computer Science, advised by Prof. Magdalena Balazinska Seattle, WA

- Research Area: Data Management Systems and Machine Learning.

Fudan University Sep 2015 – Jul 2019
BSc in Computer Science (Honors) Shanghai

- Graduated with Highest Distinction (6 out of 118).

Work Experience

Snowflake Jun 2023 – Sep 2023
Software Engineer Intern, Machine Learning Platform Team Bellevue, WA

- **Created Snowflake's Distributed Model Trainer (SLT)**: Led the design and implementation (in Python) of the first distributed PyTorch training solution within Snowflake. My work automates the model training process and creates a user-friendly interface that hides underlying complexities, enabling users to train deep models with their data inside Snowflake effortlessly.
- **Demonstrated Multi-Modal Support of SLT**: Showcased my work's support for three data modalities within Snowflake.
  - **MLP Training with Tabular Data**: Trained a Multi-Layer Perceptron using tabular data stored in Snowflake.
  - **ResNet Training with Images**: Trained an image classification model with image data stored in Snowflake.
  - **LLM Finetuning with Text**: Finetuned a large language model (stablelm-base-alpha-3b) using text data stored in Snowflake.

Microsoft Jun 2021 – Sep 2021
Research Intern, Microsoft Jim Gray Systems Lab led by Prof. Raghu Ramakrishnan Remote

- **Pioneered Tensor Query Processor (TQP)**: Led the design and implementation (in Python & C++) of the world's first query processor that compiles SQL queries into PyTorch programs and executes them on various hardware (CPUs, GPUs, TPUs, etc.).
- **Full Benchmark Support & Enhanced Performance**: Enabled full TPC-H benchmark support with TQP (until 2021 no GPU database was able to support full TPC-H benchmark), improving query execution time by 10x over specialized CPU and GPU systems (DuckDB, HeavyDB, ...) and providing acceleration for a 9x speedup for queries involving ML inference.
- **Recognized Excellence**: First-authored a VLDB paper on TQP and won the **Best Demo Award** at VLDB 2022 and **covered by Azure CTO** at Microsoft Build 2023.

Goldman Sachs Jul 2018 – Sep 2018
Summer Analyst, Engineering Hong Kong

- **Global Engineering Challenge Champion**: Clinched the Global Winner title in the Intern Engineering Challenge.
- **Revamped Critical Financial Process**: Redesigned and re-implemented the logic (in Java) for the true-up job reconciling estimated vs. actual profit and loss (PnL). Deployed enhancements led to 50% reduction in memory usage, significantly minimizing the risk of job failure.

Tencent Jan 2018 – Feb 2018
Research Intern, YouTu X-Lab led by Prof. Jiaya Jia and Prof. Yu-Wing Tai Shenzhen

- **Optimized Neural Network Inference Efficiency**: Analyzed node liveness and dependencies (in C++) in production-level deep neural networks, achieving up to 30% reduction in memory consumption through memory sharing.
- **Enhanced Data Collection and Annotation Process**: Created tools (in Python) for gathering and annotating large-scale image data, streamlining the training process for image classification models.

Selected Awards

- Madrona Prize, for the most commercializable research project, Madrona Venture & UW [GeekWire] [BusinessWire] 2022
- Best Demo Award, 48th International Conference on Very Large Databases (VLDB) 2022
- Paul G. Allen Fellowship, University of Washington 2019 – 2020
- Wangdao Scholar, Undergraduate Research Opportunities Program, Fudan University 2018
- Silver Medal, ACM International Collegiate Programming Contest (ACM-ICPC), Asia Regional 2015 – 2016
- Silver Medal (Top 100 in China), National Olympiad in Informatics (NOI), China National Finals 2014
- First Prizes, National Olympiad in Informatics in Provinces (NOIP), Guangdong Division 2009 – 2014

Selected Projects

MaskSearch: Querying Image Masks at Scale Jul 2022 – present
Project owner & leader UW

- **Developed MaskSearch**: Led the design and implementation (in Python) of a system that accelerates image retrieval queries based on mask annotations, which is essential for numerous applications such as identifying spurious correlations learned by ML models and detecting maliciously manipulated images.
- **Implemented Innovative Techniques**: Created a novel indexing technique and an efficient filter-verification query execution framework to streamline queries on mask properties.
- **Achieved Outstanding Results**: Accelerated individual queries by up to two orders of magnitude, using indexes only 5% the size of the original data, and consistently outperformed existing methods in various multi-query workloads.

Query Processing on Tensor Computation Runtimes Jun 2021 – Jun 2022
Project owner & leader Microsoft, UW

- **Pioneered Tensor Query Processor (TQP)**: Led the design and implementation (in Python & C++) of the industry's first query processor operating on PyTorch, transforming SQL queries into tensor programs.
- **Full TPC-H Support & Hardware Adaptability**: enabled TQP to support the full TPC-H benchmark on various hardware with reduced development effort, demonstrating the tensor abstraction's capability to relational SQL queries.
- **Significant Speedups**: Improved query execution time by 10x over specialized CPU and GPU systems (DuckDB, HeavyDB, ...) and realized query acceleration for a 9x speedup over CPU baselines when ML model inference is used within SQL queries.

Accelerating Queries for Neural Network Interpretation [Website] Oct 2019 – Apr 2021
Project owner & leader UW

- **Led DeepEverest Development**: Designed and implemented a state-of-the-art system (in C++ & Python) for efficiently executing interpretation queries that identify examples based on deep neural network activation patterns, by designing an efficient indexing technique and an instance-optimal query execution algorithm with critical optimizations.
- **Optimized Storage and Performance**: Accelerated individual queries by up to 63x while reducing storage requirements to less than 20% of full materialization, consistently outperforming competing baselines in various multi-query workloads that simulate DNN interpretation processes.

VisualWorld Video Data Management Project [Website] Oct 2019 – present
Project contributor UW

- **VOCAL**: a set of video data management systems that support efficient data cleaning, exploration, and organization for large-scale video data, as well as processing complex compositional queries, even when no pretrained model exists.
- **TASM**: a video storage manager which enables spatial random access to encoded videos. TASM speeds up content retrieval queries by up to 94% and 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, allowing 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 Accelerating Mobile Applications Jul 2017 – Aug 2017
Project contributor Peking University

- **Developed FPGA-Based Edge Computing Model**: Engineered a prototype (in C++ & Python) that minimizes response time and energy consumption for interactive mobile applications by offloading computation to an FPGA-based edge.
- **Proven Performance Improvements**: Achieved up to 3x/15x faster response times over CPU-based edge/cloud offloading and enhanced energy efficiency by up to 29.5%.

Publications

MaskSearch: Querying Image Masks at Scale. [Preprint] [Code] ***D. He***, *J. Zhang, M. Daum, A. Ratner, M. Balazinska.*

VOCALExplore: Pay-as-You-Go Video Data Exploration and Model Building. [Preprint] [Code] *M. Daum, E. Zhang,* ***D. He***, *S. Mussmann, B. Haynes, R. Krishna, M. Balazinska.* VLDB 2024 (to appear).

EQUI-VOCAL: Synthesizing Queries for Compositional Video Events from Limited User Interactions. [Preprint] [Code] *E. Zhang, M. Daum,* ***D. He***, *B. Haynes, R. Krishna, M. Balazinska.* VLDB 2023.

EQUI-VOCAL Demonstration: Synthesizing Video Queries from User Interactions. *E. Zhang, M. Daum,* ***D. He***, *M. Ganti, B. Haynes, R. Krishna, M. Balazinska.* VLDB 2023, Demo Track.

Query Processing on Tensor Computation Runtimes. [Paper] [MarkTechPost] [SyncedReview] [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 Tech 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] [Tech Report] [Code] [Talk] *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] [Code] [Talk] *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. [Poster] ***D. He**, Y. Chen.* CoNEXT 2018 Poster Session.

Invited Talks & Presentations

- Snowflake, Query Processing on PyTorch — Jul 2023
- UW Madison, Data Management for Model Explanation and Exploration — Apr 2023
- Huawei Cloud, Query Processing on Tensor Computation Runtimes — Feb 2023
- UW CSE Affiliates Day, Data Management for Model Exploration and Debugging — Nov 2022
- VLDB 2022, Accelerating Declarative Top-K Queries for Deep Neural Network Interpretation [Video] — Sep 2022
- VLDB 2022, Query Processing on Tensor Computation Runtimes [Video] — Sep 2022
- VLDB 2022, How Databases can Leverage the Machine Learning Ecosystem — Sep 2022
- RelationalAI, Query Processing on Tensor Computation Runtimes — Jun 2022
- Microsoft Jim Gray Systems Lab, Query Processing on Tensor Computation Runtimes — Sep 2021

Teaching & 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

Mentoring Experience

- Master / Undergraduate Students: Jason Li (2022-2023), Mona Gandhi (2022–2023), Tim Li (2022).
- Highschool Students: Parie Kumar (2022).

Professional Skills

- **Programming Languages**: C/C++, Python, Java, Pascal, Javascript, Matlab, ...
- **Machine Learning**: PyTorch, TensorFlow, Keras, Scikit-Learn, ...
- **Technical**: Database Systems (Query Optimization & Execution, Indexing Techniques), Algorithms & Data Structures, Machine Learning Systems (Software & Hardware), Video Analytics, Computer Vision, Natural Language Processing, ...
- **Other Tools**: MySQL, PostgreSQL, DuckDB, Spark, Hadoop, Hive, Google Cloud, AWS, Snowflake, Databricks, Docker, Selenium, LaTex, Git, SVN, ...