Albert Ge

http://www.albertge.com

Email : albertge@g.harvard.edu
Mobile : 408-916-8798

Education

Harvard University Cambridge, MA
M.E. Computational Science and Engineering; GPA: 4/4 *2021 - 2023*
Thesis: Workload-Aware Neural Architectures (Awarded IACS Student Thesis Scholarship)

California Institute of Technology Pasadena, CA
B.S. Computer Science; GPA: 3.8/4 *2013 - 2017*

Research Experience

Harvard DASlab - *advisor: Stratos Idreos* Cambridge, MA
Fellow of the Harvard John A. Paulson School of Engineering and Applied Sciences (SEAS) *Jan 2022 – Present*

- Developed a novel network compression method exploiting the cluster separability of feature maps at each layer, reducing the number of parameters as much as 33%. Demonstrated across a range of skip-connection based network architectures, such as ResNet and EfficientNet.
- Implemented multi-GPU support to scale feature-map identification and reduce overall compression time by 4x.

Theory of Neural Computation - *advisor: Cengiz Pehlevan* Cambridge, MA
Independent Research *Sept 2022 – Present*

- Working on exponential neural scaling laws for datasets.

Work Experience

Academia.edu San Francisco, CA
Software Engineer *2019 - 2021*

- Scaled email distribution system to over 100 million users daily, and conducted en-masse A/B tests, increasing click-through rates by 30%. Tested multi-armed bandits on recommended papers to read.
- Analyzed new user signups using Amazon Redshift and SQL. Built funnels for click-through rates on landing pages, and A/B tested new designs to increase top-of-funnel conversion.

Abbvie Stemcentrx South San Francisco, CA
Software Engineer *2017 - 2019*

- Lead developer for managing, uploading, and querying scientific pathology data. Spearheaded design of Vue.js-based front-end of an automated pipeline workflow.

Course Projects

Memory-optimized Column-Store Database Engine

- Implemented multi-threaded, shared scanning to handle batch queries, resulting in 64x speedup.
- Implemented B+trees for index point queries, reducing cache miss rate by 99% on highly selective queries
- Supported memory-conscious hash joins, including grace hash join, and demonstrated 10x speedup over nested-loop joins.

Methods of Pipeline Parallelism for Deep Learning

- Implemented a pipeline-parallel distributed training scheme for convolutional neural networks across Broadwell-CPU nodes, scaling to 1.3x speedup per additional CPU node.
- Performed roofline analysis and profiling to identify best partitioning of model layers across nodes.

Service

Graduate Advisory Committee on Diversity, Inclusion, and Leadership: Led a G1 mentoring program for students in the IACS department.

Skills

Programming: Python, C/C++, SQL, Javascript

Technologies: Pytorch, SLURM, Tensorflow, Pandas, Sklearn, React, AWS (EC2, Redshift, Kinesis), Postgres

Languages: English, Chinese