Haichen Shen

Computer Science & Engineering
University of Washington
185 Stevens Way, Seattle, WA 98195

Email: haichen@cs.washington.edu
Homepage: https://homes.cs.washington.edu/∼haichen/

Education

University of Washington — Seattle, WA
Ph.D. student, Computer Science and Engineering — Present
Advisors: Arvind Krishnamurthy
Thesis Topic: *Efficient Neural Network Execution System*

University of Washington — Seattle, WA
M.S., Computer Science and Engineering — June 2014
Advisors: David Wetherall and Aruna Balasubramanian
Thesis: *Enhancing Mobile Apps To Use Sensor Hubs Without Programmer Effort*

Tsinghua University — Beijing, China
B.Eng., Computer Science — July 2012

Interests

Deep learning system, distributed system, mobile system, and computer networks.

Projects

Nexus: Scalable and Efficient Neural Network Execution on GPU Clusters
With *Matthai Philipose and Arvind Krishnamurthy*

Nexus is a serving system for neural network inferences on GPU clusters. Nexus analyzes and schedules the DNN models at the level of linear algebra operations, which motivates several new ways to batch these operations and a batching-aware cluster resource allocation and scheduling framework to improve GPU utilization. On realistic workloads, Nexus improves efficiency by 4-19× relative to state-of-the-art baseline while staying within latency constraints >99% of the time, and scales linearly on a cluster of 64 GPUs.

TVM: End to End IR stack for AI Frameworks
With *Tianqi Chen, Thierry Moreau, Luis Ceze, Carlos Guestrin, Arvind Krishnamurthy*

TVM framework aims to bridge the gap between deep learning systems, which are optimized for productivity, and the multitude of programming, performance and efficiency constraints enforced by different types of hardware. TVM provides a common representation for deep learning computation workloads, and enables optimizations for CPUs, GPUs and other specialized hardware such as FPGAs.

Fast Video Classification via Adaptive Cascading of Deep Models
With *Seungyeop Han, Matthai Philipose, and Arvind Krishnamurthy*

This project aims to accelerate the speed of recognizing entities in every frame of video footage of day-to-day life. We demonstrate "specialized" CNNs trained for highly skewed class distributions, which is common in day-to-day videos, can be much simpler. When applied to recognizing faces in TV shows and movies we realized end-to-end classification speedups of 2.4-7.8x/2.6-11.2x (on GPU/CPU) relative to a state-of-the-art CNN, at competitive accuracy.

MCDNN: An Approximation-Based Execution Framework for Deep Stream Processing
With *Seungyeop Han, Matthai Philipose, and Arvind Krishnamurthy*

This project targets to bridge the gap between high computational demands of DNNs, and limited resource budgets such as device battery and cloud costs. MCDNN allows each request to be served approximately, by systematically trading off DNN classification accuracy for resource use. Our solution combines a system for optimizing DNNs that produces a catalog of variants of each model and a run-time that schedules these variants on devices and cloud so as to maximize accuracy while staying within resource budgets.

File Synchronization Across Multiple Untrusted Storage Services
With *Seungyeop Han, Taesoo Kim, Arvind Krishnamurthy, and Tom Anderson*
Metasync is a secure and reliable file synchronization service on top of multiple existing untrusted storage providers with no centralized server. MetaSync provides larger storage capacity, but a higher reliability and higher performance service. We devised a novel variant of Paxos, pPaxos, that provides efficient and consistent updates using unmodified APIs exported by existing services. We also built a stable deterministic replication algorithm that requires minimal reshuffling of replicated objects under service re-configuration.

Enhancing Mobile Apps To Use Sensor Hubs Without Programmer Effort
With *Aruna Balasubramanian, Anthony LaMarca, and David Wetherall*
MobileHub leverages heterogeneous hardware, such as sensor hubs, to improve the power efficiency of always-on sensing applications on mobile phones. By using information-flow tracking, we can learn how applications use sensor data, and then determines the best policy for offloading sensing tasks to sensor hubs without causing delay in the application behavior. MobileHub then rewrites the binary of a given application without programmer effort to make it more power efficient.

Publications

H. Shen, S. Han, M. Philipose, A. Krishnamurthy. *Fast Video Classification via Adaptive Cascading of Deep Models.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, July 2017. **Spotlight**.

S. Han*, **H. Shen***, M. Philipose, S. Agarwal, A. Wolman, A. Krishnamurthy. *MCDNN: An Approximation-Based Execution Framework for Deep Stream Processing Under Resource Constraints.* In Proceedings of the Annual International Conference on Mobile Systems, Applications, and Services (MobiSys), ACM, June 2016. (*equally contributed)

S. Han, **H. Shen**, T. Kim, A. Krishnamurthy, T. Anderson, D. Wetherall. *MetaSync: File Synchronization Across Multiple Untrusted Storage Services.* IEEE Internet Computing, May 2016.

H. Shen, A. Balasubramanian, A. LaMarca, D. Wetherall. *MobileHub: No Programmer Effort for Power Efficiency with Sensor Hub.* GetMobile: Mobile Computing and Communications, March 2016.

H. Shen, A. Balasubramanian, A. LaMarca, D. Wetherall. *Enhancing mobile apps to use sensor hubs without programmer effort.* In Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (Ubicomp), ACM, September 2015. **Best Paper and Gaetano Borriello Best Student Paper**.

S. Han, **H. Shen**, T. Kim, A. Krishnamurthy, T. Anderson, D. Wetherall. *MetaSync: File Synchronization Across Multiple Untrusted Storage Services.* In Proceedings of the USENIX Annual Technical Conference (USENIX ATC), July 2015.

K. Tan, **H. Shen**, J. Zhang, Y. Zhang. *Enable Flexible Spectrum Access with Spectrum Virtualization.* In Proceedings of the IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN), IEEE, October 2012.

J. Zhang, **H. Shen**, K. Tan, R. Chandra, Y. Zhang, Q. Zhang.*Frame Retransmissions Considered Harmful: Improving Spectrum Efficiency Using Micro-ACKs.* In Proceedings of the Annual International Conference on Mobile Computing and Networking (MobiCom), ACM, August 2012.

Work Experience

Google Mountain View, CA
Intern, Planet Networking Group Oct 2015 - Jan 2016
With *David Wetherall and Ashish Naik*

We developed a network monitoring system for Google WAN network. The system collects streaming data from switches and application traffic data. It detects the congestion events based on packet drop counters and filters out the non-significant events. Then the monitoring system correlates the network congestion events with application traffic to determine the culprit applications that cause the congestion, and the victims of congestion.

Microsoft Research Redmond, WA
Research Intern, Mobility and Networking Research Group June - Sept 2014
With *Matthai Philipose, Sharad Agarwal, and Alec Wolman*

We explored various techniques for compressing convolutional neural networks to reduce memory consumption and computation demand without scarifying much accuracy. We proposed the optimization that shares bottom layers among different models and applications. This project motivates the design of approximated execution framework between mobile and cloud.

Microsoft Research Asia Beijing, China
Research Intern, Wireless and Networking Research Group Apr 2010 - June 2012
With *Kun Tan and Jiansong Zhang*

We developed the software radio platform, Sora, that enables researchers to develop new wireless signal processing functions. We implemented standard wireless standards including Wi-Fi 802.11a/b/g in the development toolkit. Based on Sora, we proposed a new in-frame re-transmission scheme using symbol-level ACK to avoid re-transmit the whole packet, reducing the overhead of re-transmission. We designed an asynchronous protocols to coordinate sender and receiver using Micro-ACKs, and enable sender to re-transmit erroneous symbols immediately after the origin packets.

Programming Skills

Programming: C/C++, Python, Java, Go, Javascript, SQL, Shell, LaTeX
Deep Learning Frameworks: Caffe, Tensorflow, CNTK, MXNet

Awards

Ubicomp Best Paper Award	2015
Baidu Scholarship Award, Tsinghua University	2011
Yao Award, Tsinghua University, 2nd prize	2011
SIGCOMM Best Demo Award	2010
COSL Fellowship, Tsinghua University	2010
Honorable Mention in Math Contest in Modeling, Tsinghua University	2009
SK Fellowship, Tsinghua University	2009

Teaching Experience

Deep Learning System (UW CSE 599G1)
Tutor Spring 2017

Distributed Systems (UW CSEP 552)
Teaching Assistant Winter 2016

Network Systems (UW CSEP 561)
Teaching Assistant Fall 2013