HUI GUAN

Engineering Building II, Room 3224, 890 Oval Dr., Raleigh, NC 27695
919-649-2868 ⋄ hguan2@ncsu.edu ⋄ homepage

RESEARCH INTERESTS

My research lies in the intersection between Machine Learning and Programming Systems: Improving Machine Learning (e.g., speed, scalability, reliability) through innovations in algorithms and programming systems (e.g., compilers, runtime); Leveraging Machine Learning to improve High Performance Computing.

EDUCATION

Ph.D. in Electrical Engineering, North Carolina State University, Raleigh, USA 2013 - 2020
Advisors: Dr. Xipeng Shen, Dr. Hamid Krim Overall GPA: 4.0/4.0

B.E. in Electrical Engineering, Nanjing University of Posts and Telecommunications 2009 - 2013
Graduate with Honors, China Overall GPA: 94/100

RESEARCH EXPERIENCE

Reuse-Centric Programming System Support of Machine Learning. 2016 - present
Research Assistant *North Carolina State University*

- Proposed a flexible ensemble DNN training framework for efficiently training a heterogeneous set of DNNs; achieved up to 1.97X speedups over the state-of-the-art framework that was designed for homogeneous DNN ensemble training. (Accepted by [MLSys'20])
- Proposed in-place zero-space ECC assisted with a new training scheme, weight distribution-oriented training, to provide the first known zero space cost memory protection for CNNs. (Published in [NeurIPS'19].)
- Developed a compiler-based framework that, for the first time, enables composability-based CNN pruning by generalizing Teacher-Student Network training for pretraining common convolutional layers; achieved up to 186X speedups. (Published in [PLDI'19].)
- Accelerated CNN training by identifying and adaptively avoiding similar vector dot products during training; saved up to 69% CNN training time with no accuracy loss. (Published in [ICDE'19].)
- Improved the performance of DNN ensemble training by eliminating pipeline redundancies in preprocessing through data sharing; reduced CPU usage by 2-11X. (Published in [SC'18].)
- Accelerated K-Means configuration by promoting multi-level computation reuse across the explorations of different configurations; achieved 5-9X speedups. (Published in [ICDE'18].)
- Accelerated distance calculation-based machine learning algorithms (K-Means, KNN, etc.) by developing Triangle Inequality-based strength reduction; produced tens of times of speedups. (Published in [PLDI'17].)

NLP-based Program Optimization and Synthesis. 2016 - present
Research Assistant *North Carolina State University*

- Automatically synthesized ASTMatcher expressions based on English descriptions of to-be-found code patterns by integrating natural language dependency analysis with ASTMatcher API types and domain knowledge.
- Automatically synthesized advising tools for suggesting program optimization knowledge from HPC programming guides (CUDA, OpenCL, etc.); leveraged multiple NLP techniques including dependency parsing, semantic role labeling, TF-IDF, and topic modeling. (Published in [SC'17].)

Low-Precision SparseNN. Summer 2019
Research Intern *Facebook Inc., Menlo Park*

– Proposed several post-training 4-bit quantization approaches to reduce the memory consumption of DNN-based recommendation models (SparseNN); delivered the 4-bit quantization solution to the model serving pipeline; reduced the model size of the production Ads ranking model by 7.2X compared with the single-precision model without accuracy loss. (Accepted in [MLSys@NeurIPS'19])
– Explored both post-training quantization and quantization-aware training for quantizing computation-intensive components of SparseNN to reduce inference latency while preserving accuracy.

Data Readiness Level. 2014 - 2016
Research Assistant *North Carolina State University*

– Proposed a topological collapse-based unsupervised method for document summarization, which outperforms state-of-the-art methods on standard datasets composed of scientific papers. (Published in [SPAWC'16].)
– Proposed information-theoretic–based metrics to measure relative richness/readiness of text data to answer specific questions; validated the metrics through a text-based experiment using Twitter data.

CONFERENCES

[MLSys'20] **Hui Guan**, Laxmikant Kishor Mokadam, Xipeng Shen, Robert Patton. "FLEET: Flexible Efficient Ensemble Training for Heterogeneous Deep Neural Networks." MLSys'20. (Acceptance rate: 20.0% (34/170)).

[NeurIPS'19] **Hui Guan**, Lin Ning, Zhen Lin, Xipeng Shen, Huiyang Zhou, and Seung-Hwan Lim."In-Place Zero-Space Memory Protection for CNN." In Advances in Neural Information Processing Systems, pp. 5735-5744. 2019. (Acceptance rate: 21.2% (1428/6743))

[PLDI'19] **Hui Guan**, Xipeng Shen, and Seung-Hwan Lim. "Wootz: a compiler-based framework for fast CNN pruning via composability." In Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation, pp. 717-730. ACM, 2019. (Acceptance rate: 27.7% (76/274))

[ICDE'19] Lin Ning, **Hui Guan**, and Xipeng Shen. "Adaptive Deep Reuse: Accelerating CNN Training on the Fly." In 2019 IEEE 35th International Conference on Data Engineering (ICDE), pp. 1538-1549. IEEE, 2019. (Acceptance rate: 18%)

[SC'18] Randall Pittman, **Hui Guan**, Xipeng Shen, Seung-Hwan Lim, and Robert M. Patton. "Exploring flexible communications for streamlining DNN ensemble training pipelines." In Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis, p. 64. IEEE, 2018. (Acceptance rate: 23%)

[ICDE'18] **Hui Guan**, Yufei Ding, Xipeng Shen, and Hamid Krim. "Reuse-Centric K-Means Configuration." In 2018 IEEE 34th International Conference on Data Engineering (ICDE), pp. 1224-1227. IEEE, 2018. (short paper) (Acceptance rate: 23%)

[SC'17] **Hui Guan**, Xipeng Shen, and Hamid Krim. "Egeria: a framework for automatic synthesis of HPC advising tools through multi-layered natural language processing." In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, p. 10. ACM, 2017. (Acceptance rate: 18% (61/327))

[PLDI'17] Yufei Ding, Lin Ning, **Hui Guan**, and Xipeng Shen. "Generalizations of the Theory and Deployment of Triangular Inequality for Compiler-Based Strength Reduction". In Proceedings of the 38th ACM SIGPLAN Conference on Programming Language Design and Implementation, pp. 33-48. ACM, 2017. (Acceptance rate: 15% (47/322))

WORKSHOP & POSTER PAPERS

[MLSys@NeurIPS'19] **Hui Guan**, Andrey Malevich, Jiyan Yang, Jongsoo Park, and Hector Yuen. "Post-Training 4-bit Quantization on Embedding Tables", MLSys Workshop on Systems for ML @ NeurIPS, 2019.

[SysML'18] Yufei Ding, Lin Ning, **Hui Guan**, Xipeng Shen, Madanlal Musuvathi, Todd Mytkowicz. "TOP: A Compiler-Based Framework for Optimizing Machine Learning Algorithms through Generalized Triangle Inequality." SysML, Feb 16th, 2018, Stanford University, 2018.

[SPAWC'16] **Hui Guan**, Wen Tang, Hamid Krim, James Keiser, Andrew Rindos, and Radmila Sazdanovic. "A topological collapse for document summarization." In 2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), pp. 1-5. IEEE, 2016.

WORK UNDER SUBMISSION

[ICSE'20] Zifan Nan, **Hui Guan**, Xipeng Shen, Chunhua Liao. "Deep Semantic-Based Co-Evolvement for Synthesizing Code Analysis from Natural Language." Submitted to ICSE'20.

[NeuralNetworks] **Hui Guan**, Lin Ning, Xipeng Shen, Seung-Hwan Lim, Timothy Menzies. " A Composability-based Approach to Fast CNN Pruning." Submitted to Neural Networks.

[TPDS] **Hui Guan**, Xipeng Shen, Hamid Krim. "An Automatic Synthesizer of Advising Tools for High Performance Computing." Submitted to IEEE Transaction on Parallel and Distributed Systems.

[InformationSystems] **Hui Guan**, Yufei Ding, Xipeng Shen, Hamid Krim. "Reuse-Centric K-Means Configuration." Submitted to Information Systems.

PREPRINTS

Hui Guan, Thanos Gentimis, Hamid Krim, and James Keiser. First Study on Data Readiness Level. arXiv preprint, 2017. [PDF]

WORKING EXPERIENCE

Internship, IBM 2014 - 2016

- Built a Bluemix developer's cognitive advisor to suggest developers (target use case: node.js and Cloudant) on resources including experts, documents and repositories using graph-based querying. Crawled data from multiple sources including Stack Overflow, IBM developerWorks, Cloudant and NPM.
- Built an iOS app for displaying and interacting with information about emerging technologies. Similar technologies are recommended based on their textual descriptions using Watson services.

TEACHING EXPERIENCE

Teaching Assistant, ECE212 Fundamentals of Logic Design. 2013 - 2014

Teaching Assistant, ECE109 Introduction to Computer Systems. 2013 - 2014

AWARDS

SIGPLAN PAC Grants for PLDI'19. 2019

NSF Travel Grant for ICDE'18. 2018

IBM Ph.D. Fellowship, IBM. 2015 - 2018

Annual Outstanding Undergraduate Student Scholarship, NUPT.	2009 - 2013
National Encouragement Scholarship, Ministry of Education of the P. R. China.	2011 - 2013
National Scholarship, Ministry of Education of the P. R. China.	2010