GENG YUAN

140 Fenway R306 Boston MA 02115
yuan.geng@northeastern.edu, Google Scholar

EDUCATION

Ph.D. Candidate in Computer Engineering ***August 2017 - May 2023 (expected)***
Northeastern University, Boston, MA, GPA: 4.00/4.0
Department of Electrical and Computer Engineering
Research Advisor: Yanzhi Wang

M.S. in Electrical & Computer Engineering ***August 2014 - May 2016***
Syracuse University, Syracuse, NY GPA: 3.69/4.0
College of Electrical Engineering and Computer Science

Research Focus: *General AI Systems, Deep Learning, Efficient Training, Model Compression, DNN Acceleration and High-Performance Computing, Emerging Deep Learning Systems, Hardware-software Co-design for DNN Architectures.*

PROFESSIONAL EXPERIENCE

Snap Inc. *05/2022 - 08/2022*
Research Internship

Northeastern University *01/2020 - 05/2020*
Teaching Assistant, *Course: Advances in Deep Learning*

Mellanox Technologies, Inc. | **NVIDIA** *07/2019 - 12/2019*
Chip Design Engineer Internship

Syracuse University *08/2018 - 12/2018*
Teaching Assistant, *Course: Object Oriented Programming in C++*

Syracuse University *08/2017 - 12/2017*
Teaching Assistant, *Course: Digital Circuits Design*

Syracuse University *08/2016 - 12/2016*
Teaching Assistant, *Course: VLSI Design Methods*

** From 01/2017 to the present, all other periods not included above were covered by* ***Research Assistant*** *under the supervision of Prof. Yanzhi Wang.*

AWARDS

12/2021 The 35th Conference on Neural Information Processing Systems (NeurIPS 2021):
➢ **Spotlight Paper Award**

05/2021 Hardware Aware Efficient Training workshop in ICLR (HAET 2021):
➢ **Best Paper Award**

03/2021 The 24th Design, Automation and Test in Europe Conference (DATE, 2021):
➢ **Best Paper Nomination**

08/2020 International Symposium on Low Power Electronics and Design (ISLPED 2020) - Design Contest:
➢ **Design Contest 1st Place Winner**

06/2019 Design Automation Conference - System Design Contest (DAC-SDC 2019):
➢ **Special Service Recognition Award**

03/2018 The 19th International Symposium on Quality Electronic Design (ISQED 2018):
➢ **Best Paper Nomination**

SELECTED RESEARCH

Fast, Accurate, and Memory-economic Sparse Training on the Edge

➢ Designing an efficient sparse training framework that uses acceleration-friendly sparsity schemes, avoids dense model storage and computation, and introduces data efficiency for practical sparse training acceleration.
➢ Exploring promising directions (e.g., layer freezing, data sieving) other than increasing sparsity to effectively reduce training costs and accelerate the training process while maintaining accuracy.

Efficient Low-bit DNN Training on Edge Devices and Heterogeneous Platforms

➢ Finding a large amount of high-quality random numbers can be extracted from the neural network itself during training. The extracted random numbers can even pass the entire NIST test suite.
➢ Proposing an efficient and random number generator-free solution for the stochastic rounding in low-bit training process on hardware such as FPGA.
➢ Investigating on efficient platform-aware mixed-precision training on heterogeneous platforms.

Randomized Binary Neural Network Accelerator Design using Superconducting Technology

➢ Explore the randomized behavior of the Adiabatic Quantum-Flux-Parametron (AQFP) superconducting logic.
➢ Leveraging the randomized behavior of AQFP, develop the first AQFP-based acceleration framework for randomized BNNs and the novel accelerator architecture design.
➢ Algorithm-Architecture co-optimization to improve both model accuracy and energy efficiency, achieving a potentially 7.8×10^4 higher energy efficiency compared to the state-of-the-art CMOS counterparts.

Sanity Check for The Lottery Ticket Hypothesis

➢ Conducting in-depth explorations of the long-standing controversies in the Lottery Ticket Hypothesis (LTH).
➢ Re-define the LTH with a proposed more rigorous definition to reconcile the controversies.

High-Performance Computing Systems using DNN Model Pruning and Emerging Technologies

➢ Exploring DNN model pruning on different networks and applications.
➢ Developing model pruning for different hardware platforms.
➢ Algorithm-Architecture co-design for emerging technology-based DNN accelerators.

Front-end Design, Back-end Design, and Tapeout the DNN Inference Accelerator

➢ Design block-circulant-based DNN inference accelerator.
➢ Implement and optimize the front-end design in verilog code.
➢ Implement the back-end design and tapeout the final accelerator prototype chip.

PUBLICATION LIST (* EQUAL CONTRIBUTION)

1. [**23'DATE**] Sung-En Chang*, Geng Yuan*, Alec Lu, Mengshu Sun, Yanyu Li, Xiaolong Ma, Zhengang Li, Yanyue Xie, Minghai Qin, Xue Lin, Zhenman Fang and Yanzhi Wang, "**ESRU: Extremely Low-Bit and Hardware-Efficient Stochastic Rounding Unit Design for 8-Bit DNN Training**", in Proceedings of the 26th Design, Automation & Test in Europe Conference & Exhibition (DATE, 2023).

2. [**23'AAAI**] Zhenglun Kong, Haoyu Ma, Geng Yuan, Mengshu Sun, Yanyue Xie, and *et.al.*, "**Peeling the Onion: Hierarchical Reduction of Data Redundancy for Efficient Vision Transformer Training**", submitted and under review in the 37th AAAI Conference on Artificial Intelligence (AAAI, 2023).

3. [**23'AAAI**] Yanyu Li, Changdi Yang, Pu Zhao, Geng Yuan, Wei Niu, and *et.al.*, "**Towards Real-Time Segmentation on the Edge**", submitted and under review in the 37th AAAI Conference on Artificial Intelligence (AAAI, 2023).

4. [**22'NeurIPS**] Geng Yuan*, Yanyu Li*, Sheng Li, Zhenglun Kong, Sergey Tulyakov, Xulong Tang, Yanzhi Wang, Jian Ren, "**Layer Freezing & Data Sieving: Missing Pieces of a Generic Framework for Sparse Training**", in Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS, 2022).

5. [**22'NeurIPS**] Yanyu Li*, Geng Yuan*, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, Jian Ren, "**EfficientFormer: Vision Transformers at MobileNet Speed**", in Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS, 2022).

6. [**22'NeurIPS**] Zifeng Wang, Zheng Zhan, Yifan Gong, Geng Yuan, Wei Niu, Tong Jian, Bin Ren, Stratis Ioannidis, Yanzhi Wang, Jennifer Dy, "**SparCL: Sparse Continual Learning on the Edge**", in Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS, 2022).

7. [**22'ECCV**] Geng Yuan, Sung-En Chang, Qing Jin, Alec Lu, Yanyu Li, Yushu Wu, Zhenglun Kong, Yanyue Xie, Peiyan Dong, Minghai Qin, Xiaolong Ma, Xulong Tang, Zhenman Fang, Yanzhi Wang "**You Already Have It: A Generator-Free Low-Precision DNN Training Framework using Stochastic Rounding**", in Proceedings of the European Conference on Computer Vision (ECCV, 2022).

8. [**22'TECS Journal**] Geng Yuan, Peiyan Dong, Mengshu Sun, Wei Niu, Zhengang Li, Yuxuan Cai, Yanyu Li, Jun Liu, Weiwen Jiang, Xue Lin, Bin Ren, Xulong Tang, Yanzhi Wang, "**Mobile or FPGA? A Comprehensive Evaluation on Energy Efficiency and a Unified Optimization Framework**", in ACM Transactions on Embedded Computing Systems (TECS, 2022).

9. [**22'TODAES Journal**] Yifan Gong, Geng Yuan, Zheng Zhan, Wei Niu, Zhengang Li, Pu Zhao, Yuxuan Cai, Sijia Liu, Bin Ren, Xue Lin, Xulong Tang, Yanzhi Wang, "**Automatic Mapping of the Best-Suited DNN Pruning Schemes for Real-Time Mobile Acceleration**", in ACM Transactions on Design Automation of Electronic Systems (TODAES, 2022).

10. [**22'DATE**] Siyue Wang*, Geng Yuan*, Xiaolong Ma, Yanyu Li, Xue Lin, Bhavya Kailkhura, "**Fault-tolerant deep neural networks for processing-in-memory based autonomous edge systems**", in Proceedings of the 25th Design, Automation & Test in Europe Conference & Exhibition (DATE, 2022).

11. [**22'DAC**] Sung-En Chang*, Geng Yuan*, Alec Lu*, Mengshu Sun, Yanyu Li, Xiaolong Ma, Zhengang Li, Yanyue Xie, Minghai Qin, Xue Lin, Zhenman Fang, Yanzhi Wang, "**Hardware-efficient stochastic rounding unit design for DNN training: Late Breaking Results**", in Proceedings of the 59th ACM/IEEE Design Automation Conference (DAC, 2022).

12. [**22'ISQED**] Xiaolong Ma*, Geng Yuan*, Zhengang Li*, Yifan Gong, Tianyun Zhang, Wei Niu, Zheng Zhan, Pu Zhao, Ning Liu, Jian Tang, Xue Lin, Bin Ren, Yanzhi Wang, "**BLCR: Towards Real-time DNN Execution with Block-based Reweighted Pruning**", in Proceedings of the 23rd International Symposium on Quality Electronic Design (ISQED, 2022).

13. [**22'WWW**] Bingyao Li*, Qi Xue*, Geng Yuan*, Sheng Li, Xiaolong Ma, Yanzhi Wang, Xulong Tang, "**Optimizing Data Layout for Training Deep Neural Networks**", in Companion Proceedings of the Web Conference (WWW, 2022).

14. [**22'CLOUD**] Danlin Jia, Geng Yuan, Xue Lin, Ningfang Mi, "**A Data-Loader Tunable Knob to Shorten GPU Idleness for Distributed Deep Learning**", in Proceedings of the IEEE 15th International Conference on Cloud Computing (CLOUD, 2022).

15. [**21'ISCA**] Geng Yuan*, Payman Behnam*, Zhengang Li, Ali Shafiei, Sheng Lin, Xiaolong Ma, Hang Liu, Xuehai Qian, Mahdi Nazm Bojnordi, Yanzhi Wang, Caiwen Ding, "**FORMS: Fine-grained Polarized ReRAM-based In-situ Computation for Mixed-Signal DNN Accelerator**", in Proceedings of the 48th International Symposium on Computer Architecture (ISCA, 2021).

16. [**21'NeurIPS**] [**Spotlight (Top 3%)**] Geng Yuan*, Xiaolong Ma*, Wei Niu, Zhengang Li, Zhenglun Kong, Ning Liu, Yifan Gong, Zheng Zhan, Chaoyang He, Qing Jin, and *et.al.*, "**MEST: Accurate and Fast Memory-Economic Sparse Training Framework on the Edge**", in Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS, 2021).

17. [**21'NeurIPS**] Xiaolong Ma*, Geng Yuan*, Xuan Shen, Tianlong Chen, Xuxi Chen, Xiaohan Chen, Ning Liu, Minghai Qin, Sijia Liu, Zhangyang Wang, Yanzhi Wang, "**Sanity Checks for Lottery Tickets: Does Your Winning Ticket Really Win the Jackpot?**", in Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS, 2021).

18. [**21'ICML**] Ning Liu*, Geng Yuan*, Zhengping Che, Xuan Shen, Xiaolong Ma, Qing Jin, Jian Ren, Jian Tang, Sijia Liu, Yanzhi Wang, "**Lottery Ticket Preserves Weight Correlation: Is It Desirable or Not?**", in Proceedings of the 38th International Conference on Machine Learning (ICML, 2021).

19. [**21'DATE**] [**Best Paper Nomination**] Geng Yuan, Payman Behnam, Yuxuan Cai, Ali Shafiee, Jingyan Fu, Zhiheng Liao, Zhengang Li, Xiaolong Ma, Jieren Deng, Jinhui Wang, Mahdi Bojnordi, Yanzhi Wang, Caiwen Ding, "**TinyADC: Peripheral Circuit-aware Weight Pruning Framework for Mixed-signal DNN Accelerators**", in Proceedings of the 24th Design, Automation and Test in Europe Conference (DATE, 2021).

20. [**21'CVPR**] [**Oral Paper (Top 5%)**] Zhengang Li*, Geng Yuan*, Wei Niu, Pu Zhao, Yanyu Li, Yuxuan Cai, Xuan Shen, Zheng Zhan, Zhenglun Kong, Qing Jin, Zhiyu Chen, Sijia Liu, Kaiyuan Yang, Bin Ren, Yanzhi Wang, Xue Lin, "**NPAS: A Compiler-aware Framework of Unified Network Pruning and Architecture Search for Beyond Real-Time Mobile Acceleration**", in Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR, 2021).

21. [**21'AAAI**] Yuxuan Cai*, Geng Yuan*, Hongjia Li, Wei Niu, Yanyu Li, Xulong Tang, Bin Ren, Yanzhi Wang, "**A Compression-Compilation Co-Design Framework Towards Real-Time Object Detection on Mobile Devices**", in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI, 2021).

22. [**21'AAAI**] Yuxuan Cai*, Hongjia Li*, Geng Yuan*, Wei Niu, Yanyu Li, Xulong Tang, Bin Ren, Yanzhi Wang, "**YOLObile: Real-time object detection on mobile devices via compression-compilation co-design**", in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI, 2021).

23. [**21'ISQED**] Geng Yuan, Zhiheng Liao, Xiaolong Ma, Yuxuan Cai, Zhenglun Kong, Xuan Shen, Jingyan Fu, Zhengang Li, Chengming Zhang, Hongwu Peng, Ning Liu, Ao Ren, Jinhui Wang, Yanzhi Wang, "**Improving DNN Fault Tolerance using Weight Pruning and Differential Crossbar Mapping for ReRAM-based Edge AI**", in Proceedings of the 22th International Symposium on Quality Electronic Design (ISQED, 2021).

24. [**21'ICS**] Chengming Zhang, Geng Yuan, Wei Niu, Jiannan Tian, Sian Jin, Donglin Zhuang, Zhe Jiang, Yanzhi Wang, Bin Ren, Shuaiwen Leon Song, Dingwen Tao, "**ClickTrain: efficient and accurate end-to-end deep learning training via fine-grained architecture-preserving pruning**", in Proceedings of the ACM International Conference on Supercomputing (ICS, 2021).

25. [**21'DAC**] Pu Zhao, Geng Yuan, Yuxuan Cai, Wei Niu, Qi Liu, Wujie Wen, Bin Ren, Yanzhi Wang, Xue Lin, "**Neural Pruning Search for Real-Time Object Detection of Autonomous Vehicles**", in Proceedings of the

ACM/IEEE 58th Design Automation Conference (DAC, 2021).

26. [**21'ICLR-HAET**] [**Best Paper Award**] Xiaolong Ma*, Zhengang Li*, Geng Yuan*, Wei Niu, Bin Ren, Yanzhi Wang, Xue Lin, "**Memory-Bounded Sparse Training on the Edge**", (ICLR 2021 workshop of Hardware-Aware Efficient Training of Deep Learning Models).

27. [**21'ICCV**] Zheng Zhan, Yifan Gong, Pu Zhao, Geng Yuan, Wei Niu, Yushu Wu, and *et.al.*, "**Achieving on-Mobile Real-Time Super-Resolution with Neural Architecture and Pruning Search**", in Proceedings of the international conference on computer vision (ICCV, 2021).

28. [**21'RTAS**] Geng Yuan, Peiyan Dong, Mengshu Sun, Wei Niu, Zhengang Li, and *et.al.*, "**Work in progress: Mobile or FPGA? A comprehensive evaluation on energy efficiency and a unified optimization framework**", in Proceedings of the IEEE 27th Real-Time and Embedded Technology and Applications Symposium (RTAS, 2021).

29. [**21'RTAS**] Pu Zhao, Wei Niu, Geng Yuan, Yuxuan Cai, Hsin-Hsuan Sung, and *et.al.*, "**Industry paper: Towards real-time 3D object detection for autonomous vehicles with pruning search**", in Proceedings of the IEEE 27th Real-Time and Embedded Technology and Applications Symposium (RTAS, 2021).

30. [**21'ASP-DAC**] Hongjia Li, Geng Yuan, Wei Niu, Yuxuan Cai, Mengshu Sun, Zhengang Li, Bin Ren, Xue Lin, Yanzhi Wang, "**Real-time mobile acceleration of DNNs: From computer vision to medical applications**", in Proceedings of the 26th Asia and South Pacific Design Automation Conference (ASP-DAC, 2021).

31. [**21'TNNLS Journal**] Xiaolong Ma, Sheng Lin, Shaokai Ye, Zhezhi He, Linfeng Zhang, Geng Yuan, and *et.al.*, "**Non-Structured DNN Weight Pruning–Is It Beneficial in Any Platform?**", IEEE Transactions on Neural Networks and Learning Systems (TNNLS, 2021).

32. [**21'IJCAI demo**] Xuan Shen*, Geng Yuan*, Wei Niu, Xiaolong Ma, Jiexiong Guan, Zhengang Li, Bin Ren, Yanzhi Wang, "**Towards Fast and Accurate Multi-Person Pose Estimation on Mobile Devices**", in Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI-21) Demonstrations Track.

33. [**21'ICCAD**] Xu, Weizheng, Ashutosh Pattnaik, Geng Yuan, Yanzhi Wang, Youtao Zhang, and Xulong Tang, "**ScaleDNN: Data Movement Aware DNN Training on Multi-GPU**", in Proceedings of the 40th International Conference on Computer-Aided Design (ICCAD, 2021).

34. [**20'ISLPED**] [**Design Contest 1st Place Winner**] Geng Yuan, Wei Niu, Pu Zhao, Xue Lin, Bin Ren, and Yanzhi Wang, "**CoCoPIE: A Framework of Compression-Compilation Co-design Towards Ultra-high Energy Efficiency and Real-Time DNN Inference on Mobile Devices**", IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED, 2020).

35. [**20'ICCAD**] Gongye, Cheng, Hongjia Li, Xiang Zhang, Majid Sabbagh, Geng Yuan, Xue Lin, Thomas Wahl, and Yunsi Fei, "**New passive and active attacks on deep neural networks in medical applications**", in Proceedings of the 39th International Conference on Computer-Aided Design (ICCAD, 2020).

36. [**20'ASP-DAC**] Xiaolong Ma, Geng Yuan, Sheng Lin, Caiwen Ding, Fuxun Yu, Tao Liu, Wujie Wen, Xiang Chen, Yanzhi Wang, "**Tiny but Accurate: A Pruned, Quantized and Optimized Framework of an Ultra Efficient DNN Device**", in 25th Asia and South Pacific Design Automation Conference (ASP-DAC, 2020).

37. [**19'ISLPED**] Geng Yuan, Xiaolong Ma, Caiwen Ding, Sheng Lin, Tianyun Zhang, Zeinab S. Jalali, Yilong Zhao, Li Jiang, Sucheta Soundarajan, Yanzhi Wang, "**An Ultra-Efficient Memristor-Based DNN Framework with Structured Pruning and Quantization Using ADMM**", in IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED, 2019).

38. [**19'NANOARCH**] Xiaolong Ma, Geng Yuan, Sheng Lin, Zhengang Li, Yanzhi Wang, "**ResNet Can Be Pruned 60x: Introducing Network Purification and Unused Path Removal (P-RM) after Weight Pruning**", in 15th IEEE / ACM International Symposium on Nanoscale Architectures (NANOARCH, 2019).

39. [**18'AAAI**] Yanzhi Wang, Caiwen Ding, Zhe Li, Geng Yuan, Siyu Liao, Xiaolong Ma, Bo Yuan, Xuehai Qian, Jian Tang, Qinru Qiu, Xue Lin. "**Towards ultra-high performance and energy efficiency of deep learning systems: an algorithm-hardware co-optimization framework**", in AAAI Conference on Artificial Intelligence (AAAI, 2018).

40. [**18'ISQED**] [**Best Paper Nomination**] Xiaolong Ma, Yipeng Zhang, Geng Yuan, Ao Ren, Zhe Li, Jie Han, Jingtong Hu, Yanzhi Wang. "**An Area and Energy Efficient Design of Domain-Wall Memory-Based Deep Convolutional Neural Networks using Stochastic Computing**", in International Symposium on Quality Electronic Design (ISQED, 2018).

41. [**18'GLSVLSI**] Caiwen Ding, Ao Ren, Geng Yuan, Xiaolong Ma, Jiayu Li, Ning Liu, Bo Yuan, Yanzhi Wang. "**Structured Weight Matrices-Based Hardware Accelerators in Deep Neural Networks: FPGAs and ASICs**" in Proceedings of the 2018 on Great Lakes Symposium on VLSI (GLSVLSI, 2018).

42. [**17'MICRO**] Caiwen Ding, Siyu Liao, Yanzhi Wang, Zhe Li, Ning Liu, Youwei Zhuo, Chao Wang, Xuehai Qian, Yu Bai, Geng Yuan, and *et.al.*, "**CirCNN: accelerating and compressing deep neural networks using block-circulant weight matrices**", in Proceedings of the International Symposium on Microarchitecture (MICRO, 2017).

43. [**17'MWSCAS**] Geng Yuan, Caiwen Ding, Ruizhe Cai, Xiaolong Ma, Ziyi Zhao, Ao Ren, Bo Yuan, Yanzhi Wang. "**Memristor crossbar-based ultra-efficient next-generation baseband processors**", in IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS, 2017).

PROPOSAL IN SUBMISSION

[**NSF CSR**], "Systematically facilitating the adaptiveness and energy efficiency of DNN model training via software-hardware co-design", (PI: Prof. Xulong Tang).

COURSES STUDIED

Core Courses
Object Oriented Programming C++
Machine Intelligence / Deep Learning
Digital Machine Design
Probabilistic Methods
VLSI Design Methods

Other Courses
Computer Aided Design
Advances in Deep Learning
Digital Electronic Circuits
Data Networks: Design and Performance
VLSI Testing and Verification