XIAOLONG MA

140 Fenway R306 Boston MA 02115
(+1) 315-391-2955 ⋄ ma.xiaol@northeastern.edu, Google Scholar

EDUCATION

Ph.D. Candidate in Computer Engineering *August 2016 - present*
Northeastern University, Boston, MA, GPA: 4.00/4.0
Syracuse University, Syracuse, NY (transferred to Northeastern University in 12/2018) GPA: 3.79/4.0
M.S. in Electrical & Computer Engineering *August 2014 - May 2016*
Syracuse University, Syracuse, NY GPA: 3.67/4.0
College of Electrical Engineering and Computer Science
B.E. in Communication Engineering *September 2010 - June 2014*
Yanshan University, China
College of Information Science and Engineering

AREA OF INTEREST

1. Model compression and mobile acceleration of deep neural networks.
2. Real-time and energy-efficient deep learning and artificial intelligence systems.
3. Emerging neural network architectures and deep learning algorithms in computer vision tasks.

WORK EXPERIENCE

Northeastern University *01/2021 - present*
Research Assistant, *Advisor: Prof. Yanzhi Wang*

Alibaba Group (U.S.) Inc. *02/2020 - 12/2020*
Research Intern, *Advisor: Dr. Minghai Qin*

Northeastern University *01/2019 - 01/2020*
Research Assistant, *Advisor: Prof. Yanzhi Wang*

Syracuse University *08/2016 - 12/2018*
Teaching Assistant, *Advisor: Prof. Yanzhi Wang*

AWARDS & HONORS

Spotlight Paper Award, 35th Conference on Neural Information Processing Systems (NeurIPS), 2021.

Best Paper Award, Hardware-Aware Efficient Training (HAET) workshop in ICLR, 2021.

Contributed Article, Communications of the Association for Computing Machinery (CACM), 2021.

Best Paper Nomination, IEEE International Symposium on Quality Electronic Design (ISQED), 2018.

PUBLICATIONS LIST

Journal Publication (* equal contribution)

1. [**21'TPAMI**] Wei Niu, Zhengang Li, Xiaolong Ma, Peiyan Dong, Gang Zhou, Xuehai Qian, Xue Lin, Yanzhi Wang, Bin Ren, "GRIM: A General, Real-Time Deep Learning Inference Framework for Mobile Devices based on Fine-Grained Structured Weight Sparsity", in Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence (**Impact Factor 17.861**).

2. [**21'CACM**] [**Contributed Article**] Hui Guan, Shaoshan Liu, Xiaolong Ma, Wei Niu, Bin Ren, Xipeng Shen, Yanzhi Wang, Pu Zhao, "CoCoPIE: Enabling Real-Time AI on Off-the-Shelf Mobile Devices via Compression-Compilation Co-Design" in Proceedings of the Communications of the ACM. (**Author ordered alphabetically**)

3. [**21'TNNLS**] Xiaolong Ma, Sheng Lin, Shaokai Ye, Zhezhi He, Linfeng Zhang, Geng Yuan, Sia Huat Tan, Zhengang Li, Deliang Fan, Xuehai Qian, Xue Lin, Kaisheng Ma, Yanzhi Wang, "Non-Structured DNN Weight Pruning – Is It Beneficial in Any Platform?" in Proceedings of the IEEE Transactions on Neural Networks and Learning Systems (**Impact Factor 10.451**).

4. [**20'TNNLS**] Tianyun Zhang, Shaokai Ye, Kaiqi Zhang, Xiaolong Ma, Ning Liu, Linfeng Zhang, Jian Tang, Kaisheng Ma, Xue Lin, Makan Fardad, Yanzhi Wang, "StructADMM: A Systematic, High-Efficiency Framework of Structured Weight Pruning for DNNs" in Proceedings of the IEEE Transactions on Neural Networks and Learning Systems (**Impact Factor 10.451**).

Conference Publication (* equal contribution)

1. [**22'CVPR**] Zejiang Hou, Minghai, Qin, Fei Sun, Xiaolong Ma, Kun Yuan, Yi Xu, Yen-Kuang Chen, Rong Jin, Yuan Xie, Sun-Yuan Kung, "CHEX: CHannel EXploration for CNN Model Compression", will appear in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR, 2022).

2. [**22'ICLR**] Xiaolong Ma, Minghai Qin, Fei Sun, Zejiang Hou, Kun Yuan, Yi Xu, Yanzhi Wang, Yen-Kuang Chen, Rong Jin, Yuan Xie, "Effective Model Sparsification by Scheduled Grow-and-Prune Methods", will appear in Proceedings of the 10th International Conference on Learning Representations (ICLR, 2022).

3. [**22'ISQED - invited**] Xiaolong Ma, Geng Yuan, Zhengang Li, Wei Niu, Yifan Gong, Tianyun Zhang, Zheng Zhan, Pu Zhao, Jian Tang, Xue Lin, Bin Ren, Yanzhi Wang, "A General Pruning Framework Enabling Real-Time Inference on Resource-Limited Mobile Devices", will appear in Proceedings of the 23rd International Symposium on Quality Electronic Design (ISQED 2022).

4. [**22'FPGA**] Mengshu Sun, Zhengang Li, Alec Lu, Yanyu Li, Sung-En Chang, Xiaolong Ma, Xue Lin, Zhenman Fang, "FILM-QNN: Efficient FPGA Acceleration of Deep Neural Networks with Intra-Layer, Mixed-Precision Quantization", will appear in Proceedings of the International Symposium on Field-Programmable Gate Arrays (FPGA 2022).

5. [**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).

6. [**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, Siyue Wang, Minghai Qin, Bin Ren, Yanzhi Wang, Sijia Liu, Xue Lin, "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).

7. [**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).

8. [**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 2021) Demonstrations Track.

9. [**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).

10. [**21'DAC**] Tianyun Zhang, Xiaolong Ma, Zheng Zhan, Shaokai Ye, Kaidi Xu, Bingbing Li, Xiaolin Xu, Sijia Liu, Qinru Qiu, Makan Fardad, Xue Lin and Caiwen Ding, "A Unified DNN Pruning Weight Framework Using Reweighted Method", in Proceedings of the 58th Design Automation Conference (DAC 2021).

11. [**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).

12. [**21'DATE**] 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).

13. [**20'GLSVLSI**] Yifan Gong, Zheng Zhan, Zhengang Li, Wei Niu, Xiaolong Ma, Wenhao Wang, Bin Ren, Caiwen Ding, Xue Lin, Xiaolin Xu, Yanzhi Wang, "A Privacy-Preserving-Oriented DNN Pruning and Mobile Acceleration Framework", in Proceedings of the 20th Great Lakes Symposium on VLSI (GLSVLSI 2020).

14. [**20'PACT**] Masuma Rumi, Xiaolong Ma, Yanzhi Wang, Peng Jiang, "Accelerating Sparse CNN Inference on GPUs with Performance-Aware Weight Pruning", in Proceedings of the International Conference on Parallel Architectures and Compilation Techniques (PACT 2020, **acceptance rate: 25%**).

15. [**20'ECCV**] Xiaolong Ma, Wei Niu, Tianyun Zhang, Sijia Liu, Fu-Ming Guo, Sheng Lin, Hongjia Li, Xiang Chen, Jian Tang, Kaisheng Ma, Bin Ren, Yanzhi Wang, "An Image Enhancing Pattern-based Sparsity for Real-time Inference on Mobile Devices", in Proceedings of the 16th European Conference on Computer Vision (ECCV 2020, **acceptance rate: 27%**).

16. [**20'ICS**] Runbin Shi, Peiyan Dong, Tong Geng, Yuhao Ding, Xiaolong Ma, Martin Herbordt, Ang Li, Hayden So, and Yanzhi Wang, "CSB-RNN: A Faster-than-Realtime RNN Acceleration Framework with Compressed Structured Blocks", in the Proceeding of the International Conference on Supercomputing (ICS 2020).

17. [**20'AAAI**] Xiaolong Ma, Fuming Guo, Wei Niu, Xue Lin, Jian Tang, Kaisheng Ma, Bin Ren, Yanzhi Wang, "PCONV: The Missing but Desirable Sparsity in DNN Weight Pruning for Real-time Execution on Mobile Devices", in Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI 2020, **acceptance rate: 20.6%**).

18. [**20'AAAI**] Ning Liu, Xiaolong Ma, Zhiyuan Xu, Yanzhi Wang, Jian Tang, Jieping Ye, "AutoCompress: An Automatic DNN Structured Pruning Framework for Ultra-High Compression Rates", in Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI 2020, **acceptance rate: 20.6%**).

19. [**20'ASPLOS**] Wei Niu, Xiaolong Ma, Sheng Lin, Shihao Wang, Xuehai Qian, Xue Lin, Yanzhi Wang, Bin Ren, "PatDNN: Achieving Real-Time DNN Execution on Mobile Devices with Pattern-based Weight Pruning", in Proceedings of the 24th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2020, **acceptance rate: 18.07%**).

20. [**20'DAC**] Zhanhong Tan, Jiebo Song, Xiaolong Ma, Sia-Huat Tan, Hongyang Chen, Shaokai Ye, Yanzhi Wang, Kaisheng Ma, "PCNN: Pattern-based Fine-Grained Regular Pruning towards Optimizing CNN Accelerators", in Proceedings of the 57th Annual Design Automation Conference (DAC 2020).

21. [**20'DAC**] Chaoqun Chu, Yanzhi Wang, Yilong Zhao, Xiaolong Ma, Shaokai Ye, Yunyan Hong, Xiaoyao Liang, Yinhe Han, Yun Chen, Xiaosong Cui, and Li Jiang, "PIM-Prune: Fine-Grain DCNN pruning for Crossbar-based Process-In-Memory architecture", in Proceedings of the 57th Annual Design Automation Conference (DAC 2020).

22. [**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 Proceedings of the 25th Asia and South Pacific Design Automation Conference (ASP-DAC 2020).

23. [**20'ASP-DAC**] Xiaolong Ma, Zhe Li, Hongjia Li, Qiyuan An, Wenyao Xu, Qinru Qiu, Yanzhi Wang. "Database and Benchmark for Early-stage Malicious Activity Detection in 3D Printing", in Proceedings of the 25th Asia and South Pacific Design Automation Conference (ASP-DAC 2020).

24. [**19'ISVLSI**] Ruizhe Cai, Xiaolong Ma, Olivia Chen, Ao Ren, Ning Liu, Nobuyuki Yoshikawa, Yanzhi Wang, "IDE Development, Logic Synthesis and Buffer/Splitter Insertion Framework for Adiabatic Quantum-Flux-Parametron Superconducting Circuits", in Proceedings of the 2019 IEEE Computer Society Annual Symposium on VLSI (ISVLSI 2019).

25. [**19'GLSVLSI**] Hongjia Li, Ning Liu, Xiaolong Ma, Sheng Lin, Shaokai Ye, Tianyun Zhang, Xue Lin, Wenyao Xu, Yanzhi Wang, "ADMM-based Weight Pruning for Real-Time Deep Learning Acceleration on Mobile Devices, in Proceedings of the 2019 on Great Lakes Symposium on VLSI (GLSVLSI 2019).

26. [**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 Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED 2019).

27. [**19'NANOARCH**] Xiaolong Ma*, Geng Yuan*, Sheng Lin, Zhengang Li, Yanzhi Wang, "ResNet Can Be Pruned 60×: Introducing Network Purification and Unused Path Removal (P-RM) after Weight Pruning", in Proceedings of the 15th IEEE / ACM International Symposium on Nanoscale Architectures (NANOARCH 2019).

28. [**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 Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2018).

29. [**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).

30. [**17'MICRO**] Caiwen Ding, Siyu Liao, Yanzhi Wang, Zhe Li, Ning Liu, Youwei Zhuo, Chao Wang, Xuehai Qian, Yu Bai, Geng Yuan, Xiaolong Ma, Yipeng Zhang, Jian Tang, Qinru Qiu, Xue Lin, Bo Yuan. "CirCNN: accelerating and compressing deep neural networks using block-circulant weight matrices", in Proceedings of the International Symposium on Microarchitecture (MICRO 2017).

31. [**17'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 Proceedings of the International Symposium on Quality Electronic Design (ISQED 2017).

32. [**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 Proceedings of the 60th International Midwest Symposium on Circuits and Systems (MWSCAS 2017).

Workshop Publication

1. [**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).

2. [**20'BARC**] Xiaolong Ma, Wei Niu, Bin Ren, Yanzhi Wang, "A Desirable Sparsity Dimension for Real-time Acceleration", (Boston Area Architecture Workshop BARC 2020).

3. [**19'ODML-CDNNR**] Sheng Lin, Xiaolong Ma, Geng Yuan, Shaokai Ye, Kaisheng Ma, Yanzhi Wang, "Toward Extremely Low Bit and Lossless Accuracy in DNNs with Progressive ADMM", Workshop on On-Device Machine Learning & Compact Deep Neural Network Representations (ICML workshop 2019).

4. [**19'ODML-CDNNR**] Wei Niu, Xiaolong Ma, Yanzhi Wang, Bin Ren, "26ms Inference Time for ResNet-50: Towards Real-Time Execution of all DNNs on Smartphone", Workshop on On-Device Machine Learning & Compact Deep Neural Network Representations (ICML workshop 2019).

PROFESSIONAL ACTIVITIES

- **Reviewer:**
  IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD) 2022.
  European Conference on Computer Vision (ECCV 2022).
  IEEE Transactions on Neural Networks and Learning Systems (TNNLS) 2022.
  39th International Conference on Machine Learning (ICML 2022).
  IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2022)
  37th AAAI Conference on Artificial Intelligence (AAAI 2022).
  36th AAAI Conference on Artificial Intelligence (AAAI 2021).
  ELSEVIER Journal on Neurocomputing
  35th AAAI Conference on Artificial Intelligence (AAAI 2020).
  IEEE Transactions on Neural Networks and Learning Systems (TNNLS) 2020.
  PLOS ONE 2020.
  IEEE Transactions on Computers 2020.
  IEEE International Midwest Symposium on Circuits and Systems (MWSCAS 2019).