# XIAOLONG MA

140 Fenway R306 Boston MA 02115
(+1) 315 391 2955 ⋄ ma.xiaol@husky.neu.edu

## EDUCATION

**Ph.D. Candidate in Computer Engineering** *August 2016 - present*
Northeastern University, Boston, MA, **GPA: 4.00/4.0**
Syracuse University, Syracuse, NY **GPA: 3.79/4.0**
*Research Focus: Deep Learning, Model Compression, Computer Vision, High Performance Computing*
**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

**Top Skills**: *Python, PyTorch, TensorFlow, C++, Java, Matlab, Adobe Illustrator.*

## PROFESSIONAL EXPERIENCE

**02/2020 - now** **Alibaba Group (U.S.) Inc.** *Sunnyvale, CA*

Research Scientist Intern at Alibaba DAMO Academy.

## SELECTED PUBLICATIONS

**Journal Publication** (* equal contribution)

1. [**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, "Rethinking the Value of DNN Weight Sparsity on Hardware: Is It Truly Beneficial?" conditionally accepted in the IEEE Transactions on Neural Networks and Learning Systems (**Impact Factor 8.79**).
2. [**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" conditionally accepted in the IEEE Transactions on Neural Networks and Learning Systems (**Impact Factor 8.79**).

**Conference Publication**

1. [**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%**).
2. [**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).
3. [**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%**).
4. [**20'AAAI**] Ning Liu, Xiaolong Ma, Zhiyuan Xu, Yanzhi Wang, Jian Tang, Jieping Ye, "AutoSlim: 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%**).

5. [**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%**).

6. [**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).

7. [**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).

8. [Under review] Xiaolong Ma*, Zhengang Li*, Yifan Gong, Tianyun Zhang, Wei Niu, Zheng Zhan, Pu Zhao, Jian Tang, Xue Lin, Bin Ren, Yanzhi Wang, "BLK-REW: A Unified Block-based Pruning Framework using Reweighted Regularization Method", (submitted to **** 2020).

9. [Under review] Ning Liu*, Xiaolong Ma*, Zhengping Che, Yanzhi Wang, Jian Tang, Fachao Zhang, "Revisiting Different Pruning Schemes for DNN Model Compression", (submitted to **** 2020).

10. [Under review] Zhengang Li*, Yifan Gong*, Xiaolong Ma, Sijia Liu, Mengshu Sun, Zheng Zhan, Zhenglun Kong, Geng Yuan, Yanzhi Wang, "SS-Auto: A Single-Shot, Automatic Structured Weight Pruning Framework of DNNs with Ultra-High Efficiency", (submitted to **** 2020).

11. [Under review] Wei Niu*, Zhengang Li*, Xiaolong Ma, Peiyan Dong, Gang Zhou, Yanzhi Wang, Bin Ren, "BPDNN: A General, Real-time DNN Execution Framework on Mobile Devices with Block-based Column-Row Pruning", (submitted to **** 2020).

12. [Under review] Peng Jiang, Xiaolong Ma, Yanzhi Wang, "LAP: Locality-Aware Weight Pruning for Fast CNN Inference on GPUs", (submitted to **** 2020).

13. [Under review] 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", (submitted to **** 2020).

14. [Under review] Zheng Zhan, Yifan Gong, Zhengang Li, Wei Niu, Xiaolong Ma, Wenhao Wang, Bin Ren, Caiwen Ding, Xue Lin and Xiaolin Xu, "A Privacy-Preserving-Oriented DNN Pruning and Mobile Acceleration Framework", (submitted to **** 2020).

15. [Under review] 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" (submitted to **** 2020).

16. [**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).

17. [**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 in 25th Asia and South Pacific Design Automation Conference (ASP-DAC, 2020).

18. [**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).

19. [**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).

20. [**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", (ISLPED, 2019).

21. [**19'ICESS**] Zhe Li, Xiaolong Ma, Ji Li, Qinru Qiu, Yanzhi Wang, "Efficient Cloud Resource Management using Neuromorphic Modeling and Prediction for Virtual Machine Resource Utilization", in Proceedings of the 2019 IEEE International Conference on Embedded Software and Systems (ICESS, 2019).

22. [**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).

23. [**18'ASC**] Olivia Chen, Xiaolong Ma, Yanzhi Wang, Naoki Takeuchi, Nobuyuki Yoshikawa, "Design and Implementation of an Extremely Energy-efficient Deep Learning Accelerator Using Superconducting Logic", Applied Superconductivity Conference (ASC, 2018).

24. [**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).

25. [**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).

26. [**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).

27. [**17'ISQED**] 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, 2017). (Best Paper Nomination)

28. [**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).

**Workshop Publication**

1. [**20'BARC**] Xiaolong Ma, Wei Niu, Bin Ren, Yanzhi Wang, "A Desirable Sparsity Dimension for Real-time Acceleration", Boston Area Architecture Workshop BARC, 2020).

2. [**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).

3. [**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).

## PH.D. RESEARCH

**Graph Convolution Network (GCN) Sampling for Fast Training** *February, 2020 - Current*

· Developing algorithm for GCN layer-wise sampling.
· Designing a novel importance edge sampling method to accelerate training process.

**Reweighted Regularization Algorithm Development** *September, 2019 - Current*

· Developing the efficient DNN pruning/optimization framework by adopting Reweighted method.
· A general algorithm-sparsity co-design with high performance.
· Studying the acceleration potentials for edge computing devices of the new algorithm.

**Pattern-based Convolution Kernel Sparsity for Real-time Inference** *March, 2019 - Current*

· Discover the image enhancement properties of non-rectangular convolutions kernels in DNNs.
· Discover a new pattern-based sparsity for potential neural networks inference acceleration.
· Design the compiler-assisted acceleration framework for the new pattern-based sparsity.

**Model Quantization for High Performance Architecture** *May, 2019 - Current*

· Design multi-bit-length framework for weight, activation and gradient quantization.
· Online training with the quantized parameters.
· Implement online training on different architectures (mobilephone or FPGA).

**Model Pruning for High Performance Computing** *February, 2018 - Current*

· Implement ADMM algorithm (*PyTorch and TensorFlow*) on pruning different network structures.
· Implement both structural and non-structural pruning techniques for different objectives.
· Design specific inference system using sparse weight matrix and achieve high accuracy.
· Implement sparse matrix computing library on Android phone and get significant acceleration gain.

## SELECTED PROJECTS

**Compiler-assisted DNN acceleration framework for mobile platform (Python/Pytorch/C)**

By using the state-of-the-art pattern pruning schemes I discovered, I designed the efficient pruning algorithm and mobile acceleration framework to incorporate the generated pattern-based DNN model, and achieved real-time image inference performance.

**An extremely energy efficient in-memory computing platform (Python/Pytorch)**

Designed a highly compressed neural network model with NVM characteristic constrains. Combined channel pruning, filter pruning and weight quantization techniques with negligible accuracy loss on hardware implementation.

**IDE design for coldflux research group (Java)**

Designed a Java IDE that integrated multiple tools for our research team to edit and run different scripts on Unix system.

**ADMM-based neural networks simulator design (C++/Python/TensorFlow)**

Designed C++ and Python simulator for ADMM pruned neural network models and applied on different platforms (PC, Raspberry Pi 3, Nvidia TX1, Nvidia TX2).

**C library for Android phone (C++/Java)**

Designed a C library that can perform basic neural network computations (for both dense and sparse matrix) and implemented on an android phone to test the computation acceleration.

**Deep convolutional neural networks using stochastic computing (C++)**

I have researched on Stochastic Computing and applied this technique on deep neural networks. Stochastic Computing uses a bit-stream to represent a number which is very suitable to be applied on hardware because it has low resource requirement.

**Block-circulant matrix-based deep learning systems (Python)**

My research group has applied block-circulant matrices into deep learning systems, and could achieve reduction on weight storage and computational complexities, as well as simultaneous speedup of training and inference. This technique can be used for both software and hardware implementations, and different neural network settings.

## COURSES STUDIED

**Core Courses**
Object Oriented Programming C++
Advances in Deep Learning
Digital Machine Design
Algorithm
Machine Intelligence / Deep Learning

**Other Courses**
VLSI Testing and Verification
Computer Aided Design
Digital Signal Process
Data Networks: Design and Performance.

## PROFESSIONAL ACTIVITIES

- Reviewer:
  IEEE International Midwest Symposium on Circuits and Systems (MWSCAS) 2019.
  IEEE Transactions on Computers 2020.