Lingxiao Ma

(4th year Ph.D. Candidate)

Distributed Systems Group, Peking University, Beijing, China
xysmlx@gmail.com, xysmlx@pku.edu.cn

Research Interests

Machine Learning Systems, Graph Computing, GPU: My research works are focused on building efficient parallel systems for large-scale data analytics scenarios, e.g., deep learning, machine learning, graph computing, through leveraging modern hardware like GPU.

Education

Ph.D. in Computer Architecture — Sept. 2015 - present
Peking University, China — (Expect: Jul. 2020)
Supervisor: Prof. Yafei Dai

B.Sc. in Computer Science — Sept. 2011 - Jul. 2015
Beijing Normal University, China

Internships

Systems Research Group, Microsoft Research Asia — Nov. 2017 - present
Full-Time Research Intern, Mentor: Jilong Xue, Ming Wu — *Beijing, China*
Projects: NeuGraph (NGra), Compiler, SeerNet

Research Experience (Selected)

NeuGraph (NGra) - System for Graph Neural Networks (GNNs)
*Accepted by **USENIX ATC'19**, first author* — Dec. 2017 - present
Recent DL models have moved beyond low-dimensional regular grids to high-dimensional graph-structured data, leading to large graph-based irregular and sparse models that go beyond what existing DL frameworks are designed for. We present NeuGraph, a parallel processing framework for GNNs on top of existing DL frameworks.

- NeuGraph presents a new SAGA-NN programming model, which not only allows GNNs to be expressed intuitively, but also facilitates the mapping to an efficient dataflow representation.
- NeuGraph addresses the scalability challenge transparently through automatic graph partitioning and chunk-based streaming processing out of GPU core or over multiple GPUs.
- NeuGraph achieves efficiency through highly optimized graph operations on GPU.

Garaph - GPU-accelerated Graph Computing
*Published in **USENIX ATC'17**, first author* — Mar. 2016 - Mar. 2017
Recent advances in storage and accelerators provide the opportunity to efficiently process large-scale graphs on a single machine. Thus, we design Garaph, a GPU-accelerated graph processing system. Garaph is novel in three ways:

- Garaph proposes a vertex replication scheme to resolve GPU thread conflicts.
- Garaph adopts a balanced edge-based partition method, ensuring sequential memory access and load balance over CPU threads.
- Garaph designs a workload scheduler which considers the characteristics of processing elements and hardware.

Compiler for Deep Learning Inference
Submitted to anonymous peer-review, first author — Oct. 2018 - present
Existing deep learning (DL) frameworks leads to high and unpredictable latency for inference queries due to the coarse-grained operation and device abstractions as well as the dynamic operation-resource mappings. We propose XXX, a compiler to create a static and efficient scheduling for DL inference on GPUs.

SeerNet - Feature-Map Sparsity in Convolutional Neural Networks
*Accepted by **CVPR'19**, second author* Oct. 2018 - Nov. 2018
SeerNet is a novel and general method to accelerate convolutional neural network (CNN) inference by taking advantage of feature map sparsity. We demonstrate that a highly quantized version of the original network is suffcient in predicting the output sparsity accurately.

Publications

[1] NeuGraph: Parallel Deep Neural Network Computation on Large Graphs.
Lingxiao Ma, Zhi Yang, Youshan Miao, Jilong Xue, Ming Wu, Lidong Zhou, Yafei Dai.
2019 USENIX Annual Technical Conference (**USENIX ATC'19**) (CCF A)

[2] Garaph: Efficient GPU-accelerated Graph Processing on a Single Machine with Balanced Replication.
Lingxiao Ma, Zhi Yang, Han Chen, Jilong Xue, Yafei Dai.
2017 USENIX Annual Technical Conference (**USENIX ATC'17**) (CCF A)

[3] SeerNet: Predicting Convolutional Neural Network Feature-Map Sparsity through Low-Bit Quantization.
Shijie Cao, **Lingxiao Ma**, Wencong Xiao, Chen Zhang, Yunxin Liu, Lintao Zhang, Lanshun Nie, Zhi Yang.
30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR'19**) (CCF A)

Patents

[1] Graph Processing Method using Auto-Replication Model. 201710533444.5

Awards (Selected)

- Award for Scientific Research, Peking University 2018.12, 2017.12
- Zhitang Scholarship 2017.12
- Ph.D. President Scholarship, Peking University 2017.06
- Miaozhen Scholarship 2016.12
- Outstanding Graduate, Beijing Normal University 2015.05
- First Award, The 12th Liyun Outstanding Undergraduate Scholarship (6 of Beijing Normal University Undergraduates) 2014.12
- National Scholarship (Selected in book "Hope: Highlights of 2014 National Scholarship Winners" (ISBN: 9787301265581), 103 of 50000 Winners in China) 2014.10

Competitions (Selected)

- Silver Medal, The 39th ACM/ICPC Asia Regional Contest, Anshan Site 2014.10
- Meritorious Winner, The 30th Mathematical Contest in Modeling 2014.02
- First Prize, China Undergraduate Mathematical Contest in Modeling, Beijing Regional Contest 2013.10

Skills

Programming Language: C, C++, CUDA, Python, LaTeX, Markdown, Java, Shell
System Analysis: Performance Tuning, Outlier Diagnostics, Bottleneck Investigation
Open-source Systems: TensorFlow, PyTorch, TVM, TACO, GraphLab
Skills: GPU and Multi-Core based Programming, Graph Computing, Machine Learning, Distributed System, Data Structures and Algorithms

Other Experience

- System Administrator, Institute of NC&IS, Peking University 2016.12 - present
- Teaching Assistant, Introduction to Computing (A) 2016.9 - 2017.1

For more personal information, please visit: `https://xysmlx.github.io/`