Wenqi Jiang

wenqi.j@columbia.edu | (+41) 076 585 8978 | https://github.com/WenqiJiang

Education

Columbia University Expected Oct. 2020
M.S. in Electrical Engineering **GPA: 4.00 / 4.00**
Concentration in Data-Driven Analysis and Computation

Huazhong University of Science and Technology, China Sept. 2014 – June 2018
B.S. in Automation GPA: 3.65 / 4.00
Concentration in Pattern Recognition

Research Interest

My research interest ranges across abstraction layers in computer science from algorithm to hardware. Specifically, I have been focused on domain-specific accelerators and their deployment in datacenters. For example, I found FPGA equipped with HBM a perfect match with personalize recommendations, boosting the system performance by two order of magnitudes; coupling FPGAs with GPUs can further speedup the recommendations. In the long term, I will continually explore heterogeneous architectures related topics and hopefully conducting projects that industry can put into use.

Research Activities

MicroRec: Accelerating Deep Recommendation Systems to Microseconds Feb. 2020 – Sep. 2020

- Design and implement a high-performance personalized DNN-based recommendation system
- Resolve memory bottleneck caused by massive embedding tables lookups through hardware-software approaches
- Illustrate the usage of High-bandwidth Memory (HBM) to scale up embedding table lookups
- Revisit data structure design by applying Cartesian products to embedding tables, reducing the number of DRAM accesses
- Features a highly pipelined dataflow architecture on FPGA to eliminate recommendation latency concerns
- Achieve 168~177x speedup on embedding lookups compared to a Dell PowerEdge R630 rack server, and 21~45x speedup on end-to-end recommendation inference
- Enables low recommendation inference latency of 16.3~31.0 microseconds, almost negligible considering SLA constraints are tens of milliseconds

Towards the Gap between Sequence Model Accelerators and Real-World Applications Mar. 2019 – Aug. 2019

- Design an FPGA accelerator for language generation using Vivado HLS and deploy it on Xilinx ZCU102
- Apply a novel dataflow for matrix multiplications based on prefix-sum
- Outperform all previous sequence model FPGA accelerators on embedded platforms in terms of throughput
- Achieve a 30.02x speedup over an ARM Cortex-A53 CPU
- Point out two main factors that prevent sequence model accelerators from fully exploiting their potential
- Derive a quantitative analysis on the relationship of model size and throughput
- Discuss the DRAM access pattern of FPGA accelerators in robotics applications
- Propose the potential solution: apply embedded scalable platform consisting of multiple FPGAs

Dynamic Sampling and Selective Masking for Communication-Efficient Federated Learning Nov. 2018 – Aug. 2019

- Simulate the training process of convolutional and recurrent neural networks in federated learning by PyTorch
- First stage (finished): suppose the entire system share the same state and nodes communicate in synchronous pattern
- Apply a selective masking strategy: only transport prior updates to the central server
- Dynamic sampling leverages the communication cost and the time consumption of training process

- Reduces 80% of communication cost without losing accuracy
- Second stage (ongoing): suppose the system doesn't share the same state and communication happens asynchronously
- Propose two potential solutions to achieve near-shared state: maintaining update threshold matrix on clients asynchronously; and transport metadata, which contains statistics of the matrix, instead of transferring the entire matrix

Hercules: Analytic Engine towards Massive Structured and Unstructured Data Sep. 2019 – Dec. 2019

- A distributed database system developed by Alibaba Cloud
- Process hybrid queries containing both structured and unstructured conditions
- Apply lambda architecture with two types of indexes: graph-based index and clustering-based index
- Add new physical operators to query optimizer in order to support hybrid queries

Course Projects

Accelerate Convolutional Neural Networks by CUDA Nov. 2018 – Dec. 2018

- Take advantage of the memory hierarchy in Nvidia GPUs
- Prefetch data to overlap the computation and data transfer
- Use optimization approaches such as minimize control divergence and loop unrolling
- Construct basic building blocks, i.e., convolution layer and fully-connected layer, to ZF-Net
- Achieve up to 692% speedup compared with the naïve method

Scalable Credit Card Fraud Detection in Streaming Systems Apr. 2019 – May 2019

- Build a scalable fraud detection system using Spark Streaming
- Implement 5 algorithms (kNN, LR, SVM, ANN, and random forest) that form the Pareto curve of recall and throughput
- Train models that ensure satisfying recall on a highly unbalanced dataset by adjusting the loss function
- Dynamically select the appropriate algorithm that provides the highest recall given the current transaction throughput

Contextualized Word Vectors Nov. 2018 – Dec. 2018

- Rebuild the paper 'Learned in Translation: Contextualized Word Vectors' using Tensorflow on Google Cloud
- Train attentional neural machine translation model and generate Contextualized Vectors (CoVe)
- Compare models initialized by CoVe and GloVe on Sentiment Analysis and Question Answering datasets
- Within given training steps, models initialized by CoVe converge significantly faster than those with GloVe

Deep Reinforcement Learning for Atari Games Nov. 2018 – Dec. 2018

- Train the agents to play Breakout, one of the most popular Atari games
- Implement several deep reinforcement learning algorithms: Deep Q Learning, Deep SARSA, Double DQN, and Dueling DQN
- LeNet and VGG16 are applied as feature extraction models
- Dueling DQN with LeNet performs the best over other methods given the limited training steps

Route Tracking System for Agricultural Robots June 2017 – July 2017

- Develop a system to track the route of the agricultural robot through images
- Automatically calculate the degree of deviation of the robot
- Design a route planning algorithm to minimize the power consumption of the agricultural robot

Programming Skills

- Programming Languages: C/C++, Python, OCaml, System Verilog, Matlab, x86 Assembly
- Platforms & Frameworks: Vivado HLS, CUDA, OpenCL, OpenCV, TensorFlow, PyTorch, Keras, Spark Streaming, Apache Beam, MySQL, PostgreSQL