Heyang Qin

Department of Computer Science and Engineering
University of Nevada, Reno
1664 N Virginia St / SEM 207
Reno, NV 89557

Telephone: (775) 376-3809
Email: heyang_qin@nevada.unr.edu
https://www.cse.unr.edu/~heyangq/

Education

Ph.D. Student in Computer Science — 08/2017 - Present
Computer Science and Engineering
University of Nevada, Reno
Advisor: Dr. Feng Yan and Dr. Lei Yang

B.S. in Computer Science — 06/2017
Automation Engineering
University of Electronic Science and Technology of China

Professional Experience

Research Assistant — 08/2017- Present
Department of Computer Science and Engineering
University of Nevada, Reno

Teaching Assistant — 08/2017- Present
Department of Computer Science and Engineering
University of Nevada, Reno

Research Interests

High Performance Computing, Resource Management, Deep Learning, Reinforcement Learning, Cloud Computing, Performance Modelling and Optimization

Publication

Heyang Qin, Syed Zawad, Yanqi Zhou, Lei Yang, Dongfang Zhao, Feng Yan, Swift Machine Learning Model Serving Scheduling: A Region Based Reinforcement Learning Approach, *in Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (**SC 2019**)*, Denver, CO, USA, Nov, 2019 (Acceptance rate: 87/344=25%).

Heyang Qin, Syed Zawad, Yanqi Zhou, Sanjay Padhi, Lei Yang, and Feng Yan, Reinforcement Learning Empowered MLaaS Scheduling for Serving Intelligent Internet of Things, *IEEE Internet of Things Journal,* 2020 (Impact factor: 9.515).

Research Project

1. Region Based Reinforcement Learning Scheduling Framework for MLaaS

The parallelism settings in Machine Learning as a Service (MLaaS) have critical impact on the system performance. It is a challenge to tune the parallel config because of the complex dependency and large search space. We propose a region based reinforcement learning (RRL) approach that can converge to near-optimal config magnitude faster than the traditional reinforcement learning. The proposed RRL is prototyped and evaluated using several real-world machine learning workloads. Both theoretical analysis and experiment evaluation show that RRL outperforms state-of-the-art tuning algorithms for MLaaS.

2. RRL Plus: Adaptive Region Based Reinforcement Learning for Machine Learning

RRL is sensitive to region size. Excessive region size leads to large performance gap between RRL solution and optimal one whereas inadequate region size leads to longer learning process. We further expand the Region Based Reinforcement Learning algorithm by Bayesian optimization and heuristic algorithm and enable it to automatically adjust the region size to achieve fast converge and near-optimal solution.

3. Semi-supervised Learning for Large Scale Noisy Data

A large amount of data is available for scientific use. Unfortunately, accurate labels of training data are usually manually-labeled and expensive, leading to insufficient labeled data to train machine learning model. We solve this problem by using generative model to resemble multiple state-of-the-art models to achieve better detection from very noisy training data.

4. Efficiency-Aware Automatic Machine Learning for IoT Applications

Automatic Machine Learning(AutoML) has shown great potential in many applications. However, few studies use AutoML on resource-constrained scenarios like IoT devices. By co-designing the target model together with the backend cloud model, we are using AutoML to create a framework to automatically offload computation to the edge devices without harming the overall inference accuracy.

Service

Invited student speaker for NVASA 2019 Fall Symposium

External paper reviewer for IEEE IoT, TACO, TCC, DSN'20, SOSE'20, HPDC'19, TCC'19, BIGCOM'19, CLOUD'19, ICCCN'19, ICPE'19, SOSE'19, ICPE'18, ICCCN'18, ICMLDS'18.

Teaching

CPE201 DIGITAL DESIGN, Department of Computer Science and Engineering, University of Nevada, Reno, 08/2017 - 06/2019

ENGR100 Introduction to Engineering Design, Department of Engineering, University of Nevada, Reno, 08/2019 - 12/2019