Yu Cao

Email: yc3390@nyu.edu Phone: +1(917)575-4323
Github:https://github.com/Yucao42, Webpage:https://yucao42.github.io/

Education

New York University, Courant Institute of Mathematical Sciences Sep. 2018-Dec. 2021
Master in Computer Science

Beihang University Aug. 2011-Jul. 2018
Master and Bachelor in Instrumentation Science and Engineering

Publication and Tech Reports

- Fuqiang Zhou, **Yu Cao**, Xinming Wang, Fast and Resource-efficient Hardware Implementation of Modified Line Segment Detector. *IEEE Transactions on Circuits and Systems for Video Technology*, Co-First-Author
- **Yu Cao**, Fuqiang Zhou, Minimal Non-linear Camera Pose Estimation Method Using Lines for SLAM Applications. *IEEE Winter Conference on Applications of Computer Vision 2018*
- Cheng Tan, Zhichao Li, Jian Zhang, **Yu Cao**, Sikai Qi, Zherui Liu, Yibo Zhu, Chuanxiong GuoServing DNN models with multi-instance gpus: A case of the reconfigurable machine scheduling problem, *Arxiv 2021*
- Jianwen Jiang, **Yu Cao**, Lin Song, Shiwei Zhang, Yunkai Li, Ziyao Xu, Qian Wu, Chuang Gan, Chi Zhang, Gang Yu, Human centric spatio-temporal action localization *ActivityNet Workshop on CVPR 2018*

Experience

Research Engineer/Senior Software Engineer, Bloomberg AI Feb. 2022- present

- Collaborate with colleagues on production systems and applications.
- Design, experiment, and evaluate algorithms as well as models.
- Research and develop pricing algorithm for fix-income market.

Research Intern, ByteDance Applied Machine Learning Systems May. 2021- Aug. 2021

- Researched into scheduling algorithms(greedy and Monte Carlo Tree Search) to allocate GPU resources for DNN inference jobs to save up to 20% GPUs with Multi-Instance-GPU(MIG) feature.
- Researched into scheduling problems in pipeline DNN model serving with MIG by fitting the allocation scheduling problem into continuous optimization framework using gradient descent.

Applied Scientist Intern, AWS AI Recognition, Amazon June. 2019- Aug. 2019

- Built a multi-language scene-text localization model based on MaskRCNN detection model.
- Researched into using visual and language semantic embeddings extracted from pretrained unsupervised Auto-encoder to improve detection performance.

Algorithm Developer Intern, Face++, Megvii Inc. Nov. 2017- Jul. 2018

- CVPR 2018 ActivityNet Challenge of Spatio-temporal Action Localization(1st place)
  - Combined I3D deep learning model with Non-local module to express Actor-Target-Relationship in videos.

Teaching Assistant, Graduate Distributed Systems, New York University Aug. 2019- Dec. 2019

- Helped students on assignments of replication/Raft protocols, map-reduce, fault-tolerant key-value service.

Project

Resource Efficient Real-time Streaming Video Analytic System Oct. 2019- May. 2021

- Identified the challenge in current real-time video analytic jobs' scheduler that the processing time of frames could change because of the changing input rate or the content dependent processing logic.
- Built a distributed system(8k lines of C++) with dynamic scheduler that monitors analytic jobs' latency and throughput online and makes automatic adjustment when performance changes are detected.
- Tested on real-world gaming and traffic analysis workloads using Deep Neural Networks, our system has up to 64% more throughput(within tight latency SLO) compared with state-of-the-art system.

Awards

- McCracken Fellowship by NYU GSAS 2018-2021

Skills

- **Programming:** Python, C/C++, VHDL, Verilog, Go, Pytest Pytorch, Pandas, Linux
- **Distributed System:** Kafka, KServe, Kubernetes, Argo, Hera, Grafana, Faust, AWS S3
- **Machine Learning:** Deep learning, Random Forest, GBDT, XGBoost, Time Series
- **Quantitative Finance:** Fixed Income Bond Pricing, Backtesting, Pricing Evaluation, Municipal Bonds, Illiquid Bond Pricing, Security Similarity Modeling