Mingzhe Hu

mingzhe.hu@columbia.edu | (+1) 646-925-0794 | https://superbtum.github.io

EDUCATION

Columbia University **New York, US**
M.S. in Electrical Engineering, GPA: 3.9 / 4.0 *Expected Dec 2022*
Relevant Courses: High Performance Machine Learning, Big Data Analysis, DL, NLP, RL, ML, C++ Design

Southeast University **Nanjing, CN**
B.Eng. in Information Engineering, GPA: 3.6 / 4.0, *Jun 2020*
Exchange @ Computer Science, TUM
Relevant Courses: Computer Vision, Intro to Database, Computer Graphics Seminar

INDUSTRY EXPERIENCE

AIModelShare Initiative **New York, NY**
Software Intern, Mentor: Michael D. Parrott *Sep 2022 – Present*
- Optimized DL model replication robustness & implement plugins in neural networks
- Generated Serverless Restful API for customized machine learning models with Amazon Web Service
- Deployed tutorials of model tuning with AutoML (Dabl)

NVIDIA Corporation **New York, NY**
Software Intern, Mentor: Thomas Tang *May 2022 – Sep 2022*
- Focused on multi-camera multi-target (MTMC) analytics and set up connectivity graph for association
- Implemented Bag of Tricks (BoT) + homography matching for spatio-temporal-appearance association, and self-adaptive thresholds for intra/inter-sensor clustering, with >80% MTMC IDF1 and 5~10% improvement of single camera tracking on warehouse and retail stores
- Deployed MTMC framework into micro-batch/streaming scenario with Apache Kafka

SKILLS & HONORS

Honors: Top 1 in Megvii Workshop in Quantization with Sparsebit | Top 30% (19/77) ACM-ICPC GNY Regional | 1st place in Columbia Climate Change ML Workshop | Top 3% in Tianchi SVHN Detection & Recognition Challenge | Top 5 in Megvii Workshop in Mobile AI Photography of RAW Image Denoising
Programming Languages: Python, CUDA, C/C++, Cython, Matlab
Platforms and Tools: Google Cloud Platform, Amazon Web Service, Airflow, Twitter Streaming, Apache Spark/Kafka
Packages: PyCUDA, PyOpenCL, PyTorch, Tensorflow, OpenMP, OpenAI Gym, PyMySQL, TensorRT, Rapids
Operating Systems: Windows, Linux (CentOS, Ubuntu, Debian)

PROJECTS

Unsupervised Object Reidentification in Smart Intersections **New York, NY**
Research Assistant, Kostić Lab in COSMOS Smart Traffic with Prof. Zoran Kostić *Jan 2022 – May 2022*
- Collected and annotated YoloV4-detected pedestrian & vehicles in intersection in overlapped & sequential, daytime & nighttime scenarios with ≥1500 instances and ≥100 identities
- Designed ResNet-IBN and ViT backbone under contrastive learning logic with gradient accumulation, cluster-level memory update and DBSCAN/Infomap for pseudo-label generation
- Evaluated with domain adaption and improved mAP to 89% (pedestrian) & 79% (vehicle) with bag of tricks

Transformer Quantization on Text Classification Task **New York, NY**
Project Leader, Attached Course: Deep Learning in System Performance *Jan 2022 – May 2022*
- Fully quantized 8-bit transformer with exponential moving average observer to compress run-time model
- Achieved close text classification accuracy with AGNews dataset while reducing model size to 30%
- Explored impact of pre-trained weight and quantization latency, with 5% accuracy increase

Real-time New York Traffic Heatmap **New York, NY**
Project Leader, Attached Course: Large Scale Stream Processing *Jan 2022 – May 2022*
- Fetched real-time road condition and traffic density every 15 minutes with Spark Streaming
- Preprocessed RDD data with spark filtering, duplication removal and K-Means to reduce size to 10%
- Accelerated from 10 min to 20 sec with asynchronous REST API request, keep-alive, and FAIR scheduler

Acceleration of GloVe Representation on Heterogenous Platform **New York, NY**
Project Leader, Attached Course: Heterogenous Computing *Sep 2021 – Dec 2021*
- Worked with handcrafted CUDA and designed work-efficient sum and maximum finder, with 5 times faster
- Created Bag of Tricks (BoT) for GPU arrays alignment and efficient atomic addition logic
- Achieved ≥ 80 times faster than Numpy in naïve version and comparable speed with PyTorch