QIANHAO ZHANG

✉ zhangqianhao1408@gmail.com · ☎ (+1) 510-993-4360 · in qianhaoz · jasonlovescoding

EDUCATION

Carnegie Mellon University - School of Computer Science Pittsburgh, PA

M.S. in Computer Vision | Current GPA: 4.22/4.3 Dec. 2020

Beihang University - School of Computer Science and Engineering Beijing, China

B.Eng. in Computer Science and Technology | GPA: 3.78/4, Graduation with Honors Jul. 2019

University of Toronto - Faculty of Applied Science and Engineering Toronto, ON

Scholarship-Funded Exchange Program | GPA: 3.88/4 Dec. 2017

PROFESSIONAL EXPERIENCES

Nuro Inc. Mountain View, CA

Senior Software Engineer, Perception | Python, C++ Jun. 2024 - Present

Develop state-of-the-art perception models & implement associated train/deploy infrasturcture

- Perform in-training and post-training profiling to identify the latency/memory/tech-debt bottlenecks
- Reduced **10%** latency by re-modeling the anchor-based design with heatmap-based design and re-implementing corresponding custom dynamic ops with native static ops
- For custom ops that can't be replaced, implemented their FP16 counterparts to achieve **2x speedup**
- Optimized **20%** memory with attention-based detection model to fully exploit the feature sparsity
- Re-developed detection model with basic tf/keras3/torch ops, meanwhile ensuring the whole model's numerical & speed parity across backends, as part of the team effort to move away from tensorflow towards pytorch

Software Engineer, Perception | Python, C++ Jan. 2021 - Jun. 2024

Develop modernized perception modeling infrastructure

- Developed keras-based MMDet-like framework that **unified the modeling workflow for perception team**
- Implemented unified TF-TRT custom operator API that supports automatic TensorRT export with custom TF ops
- Re-implemented the entire camera-lidar 3D detection workflow (from data generation, model training to final deployment) with frameworks above to showcase its better performance (**significantly improved APs** with joint temporal training) and debuggability (ultimately getting **5x higher hours per interruption** without NaN / OOM, etc.)

SenseTime Co., Ltd. Beijing, China & San Jose, CA

Research Intern (San Jose Office) | C++, Bash, Python May. 2020 - Aug. 2020

Compression and quantization of neural networks for camera-related CV tasks on smartphones

- **50%** channel-pruning compression of CNN to obtain fine-grained quad bayer captured by 2x2 on-chip lens), enhanced the light-weight model (Python) for low-exposure frames with hard example fine-tuning
- **5x speedup** of CNN for bayer demosaicking on Xiaomi phone's raw data, achieved by mixed-bitwidth (16-bit activation and 8-bit weight) quantization-aware training (Python) with AIMET toolbox
- Developed the deployment pipeline for CNN models on smartphones (C++ and bash scripts), verified the model performance on the **DSP/CPU** of an Oppo Reno 2 and a google Pixel 3 with SNPE toolchain

Research Intern (Beijing Office) | C, C++, Python Feb. 2018 - Jul. 2019

Performance optimization and pipeline automation for deep learning frameworks and packages

- Developed pytorch-onnx-caffe conversion and profiling package supporting **all neural network layers**, effectively bridged the gap between research teams (training) and engineering teams (deployment)
- Designed easy-to-use, modularized APIs that successfully worked with models within a wide variety such as pedestrian re-ID, face verification, car detection, etc. (number of users soon **exceeded 300** since first release in a month)
- Implemented novel neural network layers (time-shift operation, correlation convolution, etc.) in Caffe (C++) with research teams, **halved the train-test-deploy response cycle of any new model**
- Developed inference framework (C) optimized for x86 processors with MKL-DNN, **2x speedup** compared to regular Caffe, used as deployment framework on development boards and light-weight chips

Robotics Institute, Carnegie Mellon University Pittsburgh, PA

Student Researcher, supervised by Prof. John Galeotti | Python, C++ Feb. 2020 - May. 2020

Develop stateless relocalization module to fight the drifting problem in long-range UAV flights

- Implemented a fully convolutional neural network for scene coordinate regression, and applied **differentiable RANSAC with PnP algorithm** on scene coordinates for pose estimation
- Leveraged GPS and structure-from-known-motion with OpenMVG to obtain **high-quality ground truth** for training
- Averagely **<3m, <0.3° error** tested on 10-kilometer flight data, **<1m, <0.1° error** tested on 2-kilometer flight data

FHL Vive Center for Enhanced Reality, UC Berkeley Berkeley, CA

Student Assistant III, supervised by Dr. Allen Yang | Python, C++ Jun. 2019 - Sept. 2019

Develop and review new features for OpenARK

- Implemented **ICP algorithm** for SLAM module, stabilized the trajectory on texture-sparse frames
- Implemented a Caffe-based web demo for human face registration & verification

Visiting Student Researcher, supervised by Dr. Allen Yang | Python, C++ Jul. 2018 - Oct. 2018

Design a loop closure detection module and improve localization module for the lost track problem in VR/AR scenarios

- Designed a **feature-pyramid siamese network** for loop closure detection w/ comparable performance to ORB-SLAM
- Synthesized a **large-scale** (∼150,000 images) indoor environment dataset with Unity3D and SunCG for train & test

Awards and Certificates

An Image Retrieval System Based on Natural Language Captioning, CN Patent Aug. 2019

• Automatic image captioning upon uploading, used BLEU score as the key for retrieval, enabling descriptive search

1st**-place Winner** with ¥10,000 (∼$1,500) Prize, BeyondSoft Tech Challenge on Motion Evaluation Nov. 2018

• Designed neural network to evaluate motion quality for athletes / rehabilitating patients on inertial data

National Scholarship for Academic Excellency, Chinese Ministry of Education Nov. 2017

• Top-level scholarship awarded nationally to recommended students for their academic excellency

Skills

Python, C, C++, Bash; Pytorch, Tensorflow, Keras, SciKits; TensorRT, ONNX, OpenCV, OpenMVG