# Rocking Lai

aa155495@gmail.com • +886-919-240478 • Taipei, Taiwan

## Introduction

I am a deep learning engineer as well as professional software engineer. I take part in various deep learning & computer vision projects, from software application to AI chip. Also have the experience of developing video surveillance system (full-stack). In addition, I strive for learning and constantly looking for ways to share my knowledge. I am the mentor of Coursera, which helps students in learning computer vision and machine learning.

## Skills

### Domain Knowledge

GPU Programming | Deep Learning Compiler | Neural Network Quantization | Computer Vision Algorithm | Surveillance System

### Tool

C++ | Python | Pytorch | OpenCV

## Experience

1

### **Machine Learning Software Engineer**, AMD, Sep 2021 - Present

- Develop GPU backend of various deep learning framework. Eg. PyTorch, Tensorflow, Onnx runtime
  ***a.*** Implement and optimize GPU kernel
  ***b.*** Proposed complex fused kernel, such as gemm + layernorm, gemm + reduction, gemm + softmax.
  ***c.*** Low precision arithmetic for deep learning.

2

### **Deep Learning Engineer**, CVITEK, Oct 2019 - Sep 2021

- Spin off from **Bitmain**.
- Mainly focus on deep learning accelerator (so called NPU or TPU) and stereo vision accelerator.
- Deep learning accelerator
  ***a.*** Research and prototype the int8, int4 and bf16 quantization algorithm.
  ***b.*** Mix-precision algorithm based on Post training, result had applied patent in US and CN (US20220129736A1, CN114492721A).
  ***c.*** Model compression flow (heterogeneous quantization).
  ***d.*** Design quantization flow in deep learning compiler (graph optimization, calibration, fine tuning, mix-precision).
  ***e.*** Implement frontend of deep learning compiler (high level optimization and lowering to low level IR).

*f.* Inference simulator via high level IR.
*g.* Co-work with IC designer to design the cmodel of AI accelerator.

- Stereo Vision accelerator
  *a.* Prototype the hardware-friendly stereo matching algorithm.
  *b.* Co-work with IC designer to design the cmodel.

3 **Deep Learning Engineer**, Bitmain, Aug 2018 - Oct 2019

- Mainly focus on deep learning accelerator (so called NPU or TPU and so on)
- Research of deep learning algorithm for edge AI accelerator.
  *a.* int8 and bf16 quantization algorithm.
  *b.* Post training based mix-precision algorithm.
  *c.* Quantization-aware training flow.
- Deep learning compiler
  *a.* Quantization tool (calibration, fine tuning).
  *b.* Inference simulator for high level IR (cpu & gpu).
  *c.* Co-work with IC designer to design the cmodel of AI accelerator.

4 **Algorithm Engineer**, ULSee, Jun 2017 ~ Aug 2018

- Vision algorithm for robot (object detection, gesture recognition, posture recognition).
- Facial landmark tracking algorithm.
- Improve face recognition flow.
- Driver fatigue detection, phone talking detection for ADAS (Advance driver assistance system).
- Plan & design the face recognition system for various projects. (IP camera integration, video management, lead the scrum flow).

5 **Software Engineer**, NUUO, Sep 2014 ~ Oct 2017

- Design and maintained the Network video recorder (NVR). It is an embedded Linux, which can received video stream from IP camera, various type of recording, video analytic, third party integration.
- Develop new features for NVR client, which can play live video, playback video, smart search video, event management...etc.
- Develop SDK of NVR, provide a way for third party to integrate our NVR.

## Volunteer

Course Mentor, Jan 2017 - Feb 2019

- Machine Learning, offered by Stanford University.
- Fundamentals of Digital Image and Video Processing, offered by Northwestern University.

## Education

Master's degree, Computer Science and Information Engineering

National Chiao Tung University (2012 - 2014)

Bachelor's degree, Computer Science

National Chiao Tung University (2008 - 2012)

## Publication

- **Toward Community Sensing of Road Anomalies Using Monocular Vision**, IEEE Sensors Journal (Volume:16 , Issue: 8 ) 2016
- **Vision-Based Road Bump Detection Using a Front-Mounted Car Camcorder**, IEEE International Conference on Pattern Recognition (ICPR 2014)

## Patent

- **Mixed-precision quantization method for neural network**, US20220129736A1 (2021)
- , CN114492721A (2020)