Prathamesh Mandke

+1(540) 252 9660 • mandkepk97@gmail.com • pkmandke.github.io
pkmandke • San Diego, CA

Education

Virginia Polytechnic Institute and State University — **Blacksburg, VA**
Master's in Computer Engineering, GPA: 4.0/4.0 — *Aug 2019 – May 2021*
Specialization: Software & Machine Intelligence

College of Engineering, Pune — **Pune, India**
B.Tech Electronics & Telecommunication GPA: 9.11/10 — *Aug 2015 – May 2019*
Minor in Computer Engineering

Skills

PyTorch, Deep Learning, Model Efficiency, ML Infra, Python, C++, Linux, Git

Experience

Qualcomm, Inc. — **San Diego, CA**
Senior Machine Learning Engineer — *Jul 2021 – Present*

- Implemented Qualcomm's first LLM LoRA fine-tuning on SnapDragon mobile CPU using PyTorch in C++.
- Leveraged LLM block-quantization, checkpointing, etc. to significantly reduce peak memory, US patent accepted.
- Worked with MSFT Research on co-developing a Federated Learning SDK using C++/gRPC/Azure. Work published at Interspeech'23: `https://www.isca-archive.org/interspeech_2023/mandke23_interspeech.pdf`
- Played key role in developing end-to-end SW for Federated and On-Device Personalization R&D using C++/Python. Work demonstrated at NeurIPS'21 (`https://youtu.be/fTVmDzFKjqI`) and NeurIPS'23(`https://youtu.be/jbi9Ig1wPwc`).
- **Skills/Tools**: PyTorch, ML Infra, Python, C++, gRPC, LLM/LVM

Virginia Tech — **Blacksburg, VA**
Graduate Research Assistant — *Sept 2020 – May 2021*

- **Domain**: Deep Learning for Scientific Applications (NSF Research Grant)
- Worked with Prof. Anuj Karpatne on weakly supervised semantic segmentation.
- Developed PyTorch based training and inference SW across multi-GPUs for GANs, auto-encoders, etc.

Qualcomm, Inc. — **(Remote) San Diego, CA**
Internship — *Summer 2020*

- Worked with the AI Model Efficiency Toolkit (AIMET) team: `https://github.com/quic/aimet`
- Developed visualization utilities for Model Quantization ops in graph & integrated with Netron using ONNX.
- Implemented a generic utility to convert AIMET's internal IR to ONNX.

Flytbase, Inc. — **Pune, India**
Internship — *Summer 2019*

- Worked on ML based barcode localization for drone based warehouse automation applications.
- Trained Yolo, Faster RCNN and SSD models on NVIDIA GPUs using TensorFlow/Keras.
- Explored embedded deployment of models on the Intel Neural Compute Stick.

Projects

Deep Knowledge Transfer: CNN Model Compression for OpenCL-FPGA deployment — **Dec'18 - May'19**

- Bachelor's Thesis: Implemented knowledge distillation (KD) in FaceNet CNN for model compression using TensorFlow.
- Trained MobileNet CNN using 1M VGG dataset using KD to achieve 80% accuracy with a 75% smaller model footprint.
- Explored embedded deployment on Intel's DE10 Nano FPGA SoC. Details: `https://bit.ly/github-kd-cnnt`.

Human Posture Recognition using Artificial Neural Networks — **Feb'18 - May'18**

- Designed and developed an end-to-end system to classify human postures on a Raspberry-Pi (R-Pi) using ML.
- Designed and built PCB node with IMU sensor and HTTP/TCP SW for comms with central R-Pi server.
- Trained and deployed ML model on R-Pi implemented using pure numpy to obtain classification accuracy of 97.5%.
- Code: `https://bit.ly/20vel6a`. Dataset: [github].

[Open Source] Arbitrary Layer CAMs in PyTorch based CNN models — **Nov'20**

- Added feature to compute Class Activation Maps of any arbitrary layer in a PyTorch based CNN model graph.
- Pull Request reviewed and merged in open source repository: `https://github.com/frgfm/torch-cam/pull/21`

Lempel-Ziv-Welch Text File Compression - A python package — **Apr'18 - Sept'18**

- Designed and implemented a python package for UTF-8 file compression
- Achieved compression ratio 50% along with O(log(n)) phrase look-up time. GitHub: `https://bit.ly/2vhGPHy`
- Studied compression ratio as a function of different file probability distributions.

Awards & Publications

- List of patents and publications: `https://scholar.google.com/citations?user=wB6ZbbMAAAAJ`
- Recipient of the prestigious **Narotam Sekhsaria Scholarship** for graduate school. `https://pg.nsfoundation.co.in/`