Jeremy Malloch

jeremalloch@gmail.com
github.com/jeremalloch
malloch.ca

Skills

Python, C++, PyTorch, ML Modelling, ML Deployment

Work Experience

Machine Learning Engineer – Cruise *San Francisco, CA* **July 2021 – Present**

- Unblocked Cruise's return to driverless operation by leading Unified Tracker V2 model release that resolved critical ultra-nearfield risks (largest perception risk area)
    - Worked with upstream detection teams to:
        - Root cause false positive detections
        - Mine & label targeted data
        - Jointly train detection models & tracking model with new data
    - Model shipped within 4 weeks of the Unified Tracker V1 release, ahead of schedule, unblocking driverless operation by end of 2024
    - Saw net 13 major progressions in Ultra Nearfield end to end suite
- Contributed multiple modelling improvements to novel Unified Tracker V1 multi-modal, temporal tracking transformer model:
    - Increased model flop utilization by **32%**
    - Stabilized training by analyzing model activation magnitudes
    - 20% reduction in velocity flips
    - Created model attention metrics, directly leading to modeling improvements:
    - 11% reduction in centroid error, 12% reduction in velocity error, 3.25x as many major progressions vs regressions
    - Increased recall of pedestrians from **95.03% to 98.33%**, with improved F1 scores in 5/6 semantic classes
    - Led deployment, contributed C++ changes for 3x speedup in on car node, coordinated with downstream consumers to resolve latency blockers
- Accelerated inference of multiple deployed models by developing in-house PyTorch inference optimization library (quantization, pruning) with **torch.fx**
- Led partnership to build compiler frontend / transformation for bit-accurate emulation of DNN inference on future custom inference hardware
- Collaborated with open source, contributing bug fixes and reporting issues to TensorRT, PyTorch, and Netron

Autonomous Systems Software Intern - Apple *Sunnyvale, CA* **Jan 2019 – Aug 2019**

- Contributed to machine learning compiler/inference framework in **C++ & Python**
- Researched & prototyped DNN quantization & sparsification pipeline, resulting in up to **40% faster** runtime of deployed models

Education

University of Waterloo **Waterloo, ON**

BMath in Computer Science and Combinatorics & Optimization *2021*

Relevant Courses: Algorithms, Operating Systems, Real Analysis, Group Theory, Mathematical Optimization (Convex & Linear Programming), Deep Learning in Discrete Optimization (graduate)