Benedek Hegedus

b.hegedus45@gmail.com |+1 778 229 6240

Vancouver, BC

https://www.linkedin.com/in/benedek-hegedus

Portfolio: https://www.benihegedus.com

Languages**: Python, C++, Assembly, SysVerilog**

Research interests**: Active Inference, Spectral Graph Theory**

EDUCATION

The University of British Columbia Vancouver, BC

Bachelor of Applied Science in Integrated Engineering *Sep* 2016 – Dec 2021

Specialized in Computer and Electrical Engineering

EXPERIENCE

Huawei Technologies Vancouver, Canada

Machine Learning Engineer – AI Accelerator, Self Driving ***(Python, C++)*** Jan 2022 – Jan 2023

- Re-implemented and converted models from PyTorch through ONNX to run on specialized AI accelerator hardware. Conducted detailed runtime profiling analysis: inference latency, compute hardware usage vs data movement bottlenecks. Models include autonomous driving SOTA – PointPillars, Lift-Splat-Shoot, etc.
- Optimized models and operators to utilize AI hardware to the fullest, resulting in massive reduction in inference time, up to 90%.
- Designed improved network architectures by hardware-aware design, increasing performance (relative 12%) of the models without compromising inference time – Image/LIDAR sensor fusion, perception, planning. Also built a codebase to easily do experiments and export hardware accelerated version of the model.
- Worked across the whole autonomous driving AI stack, integrating, and optimizing (inference time) multiple models in the pipeline and leading on-device deployment.
- Independently developed scripts and tools to automate model conversion, quantization, evaluation, and analysis steps, significantly increasing robustness and productivity of the end-to-end process, saving the team hours daily.
- Conducted competitor technical analysis in self driving space to evaluate different technical directions, future trends and provide valuable insights to multiple sub-teams, shaping the research direction.

Huawei Technologies Vancouver, Canada

AI researcher Co-op in Computer Vision ***(Python, C++)*** Jan 2020 – September 2020

- Converted models from TensorFlow and PyTorch to run on Atlas200DK (AI accelerator) board by using equivalent models with different operators. Models include OpenPose based keypoint detection and Transformer based language model. Applications include Signlanguage Translation, Fall Detection. Built the entire on-device pipeline as the sole developer.
- Created Hand Gesture Controlled RC Car opensource project to showcase hardware connections with Atlas200DK. This independent project was the first hardware-based project and formed the basis for many future projects.
- Implement Python based Atlas200DK projects in C++ to optimize inference, pre-processing, and post-processing time.

Laser Zentrum Hannover e.V Hannover, Germany

Machine learning ***(Python)*** *– intern* May 2019 – Dec 2019

- Built a dynamic data acquisition and camera calibration program that fully automated the data collection process. This was a significant improvement as the data was previously collected manually.
- Integrated the data acquisition system with a live post-processing algorithm. This reduced the size of saved frames from 4mb to 2kb while maintaining useful information. It worked by cropping frames around the ROI, which was computed from positions of the laser.
- Used PyTorch and Keras to create neural networks for classification.
- Implemented a custom Recurrent-CNN in PyTorch (for video classification) and achieved a classification accuracy (4 classes) of 77%. The previous best was 37%.