Benedek Hegedus

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

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 hardware **(Python, C++)*** Jan 2022 – Jan 2023

- ♦ 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.
- ♦ Optimized models and operators to utilize AI hardware to the fullest, resulting in massive reduction in inference time.
- ♦ Co-designed SOTA network architectures that are optimized for utilizing AI accelerator hardware.
- ♦ Worked across the whole autonomous driving AI stack, integrating and optimizing multiple models in the pipeline.
- ♦ Developed scripts and tools to automate model conversion, quantization, evaluation and analysis steps, significantly increasing robustness and productivity.
- ♦ Conducted competitor technical analysis in self driving space to evaluate different technical directions, future trends and provide valuable insights to the team.

Huawei Technologies Vancouver, Canada

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

- ♦ Convert 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.
- ♦ Create Hand Gesture Controlled RC Car open source project to showcase hardware connections with Atlas200DK.
- ♦ Pipeline multiple deep learning models to implemented embedded version of computer vision application.
- ♦ Review SOTA research papers in Computer Vision and AI to understand trends in model architectures.
- ♦ 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%.