Jeongwon Her

E-mail: jwher96@snu.ac.kr
Blog: https://jwher.github.io/posts
+8210-8520-3971

PROFESSIONAL EXPERIENCE

Research on Inference Optimization for AI Models (Seoul Nat'l Univ.) Sep. 2023 – Present

- Developed quantization-aware optimization techniques to improve throughput of low-precision (INT8) inference on deep learning accelerators (NPU).
- Minimized accuracy degradation by replacing PyTorch-based CV model operators and applying Knowledge Distillation and Quantization-Aware Training (QAT)
- Verified techniques to restore accuracy degradation on NPU by updating tensor scales
- Proposed a mixed-precision and pipeline parameter search methodology to optimize operator-device mapping on heterogeneous accelerators (GPU, NPU)
- Optimized 11 vision models on Jetson AGX Orin with TensorRT, yielding up to 1283% speedup and 77% energy savings over baseline implementations without optimization.
- Developed inference engines for Telechips NPU and Jetson AGX Orin; deployed models on embedded systems
- Released research results as an open-source project on GitHub

MLOps Platform Development and Operation (Upstage) Sep. 2022 – Aug. 2023

- Designed object storage system using MinIO (S3 compatible) with fine-grained data access control (Django)
- Developed Python-based API server and integrated open-source ML serving libraries (Kubernetes, BentoML)
- Managed GPU Farm infrastructure with Kubernetes and ArgoCD; built deployment/testing environments using minikube and skaffold
- Operated distributed training pipelines in GPU farm environments

Automated Training Pipeline Development (SNUAILAB) Sep. 2021 – Aug. 2022

- Built a semi-supervised training pipeline tailored for real-world traffic surveillance (VMS system)
- Implemented inference/training APIs using FastAPI; designed data storage with MongoDB and PostgreSQL
- Orchestrated batch workflows using Apache Airflow; managed models and logs with MLflow

EDUCATION

Seoul National University Sep. 2023 – Aug. 2025 (Expected)
M.S. in Computer Science and Engineering

Dongguk University Mar. 2015 – Feb. 2021
B.S. in Computer Science and Engineering
GPA: 3.80/4.5 (Major GPA: 4.02/4.5)

PUBLICATIONS

- Jeongwon Her et al., *Real-Time 3D Object Detection Using N-Dolphin Embedded NPU*, Korea Computer Congress (KCC), 2024.
- Jeongwon Her et al., *Computer Vision Application Optimization Methodology on NVIDIA Jetson Boards with TensorRT*, submitted to ACM Transactions on Architecture and Code Optimization (TACO), under review.

SKILLS

Languages: Python, Java, C++, OpenCL/CUDA
Frameworks: PyTorch, TensorFlow, TensorRT

LEADERSHIP & COMMUNICATION EXPERIENCE

Deep Learning Paper Reading Group (YouTube) Sep. 2022 – Present

- Selected and reviewed state-of-the-art Vision/Language papers every 3 weeks; produced summary videos for YouTube

Ecological Education Program (Cambodia) Jul. 2023 – Present (1 week/year)

- Delivered environmental education to middle school students in Cambodia as part of an international outreach program

Samsung AI Expert Program (Teaching Assistant) Aug. 2024 – Dec. 2024

- Conducted practical lectures on TensorFlow/ResNet model pruning and lightweighting
- Advised engineers on model pruning and quantization strategies for on-device temperature calibration in Galaxy RF modules

Military Service, ROK Army Feb. 2016 – Nov. 2017

- Served as squad leader.