Pang-Chun, Chung

+886 966 368 917 | bonginn0908@gmail.com | LinkedIn | GitHub | Blog | Hsinchu, Taiwan

Education

National Yang Ming Chiao Tung University — Hsinchu, Taiwan
M.S. in Computer Science and Engineering(Incoming) — *Starting Sep 2026*

National Yang Ming Chiao Tung University — Hsinchu, Taiwan
B.S. in Computer Science; ***GPA: 3.84/4.30*** — *Sep 2022 – Expected Jun 2026*

- **Relevant Coursework:** Deep Learning, Machine Learning, Edge AI, Compiler Design, Operating Systems, Computer Organization, Data Structures, Algorithms

Research Experience

NYCU EdgeAI Lab (Advisor: Prof. Kai-Chiang Wu) — Hsinchu, Taiwan
Undergraduate Researcher — NSTC Undergraduate Research Project Fellowship — *Jul 2025 – Present*

- Deployed and benchmarked **LLaMA-2-7B** inference on **Jetson Orin NX** using **TensorRT-LLM**, evaluating quantized configurations under tight memory, power, and latency constraints.
- Built a benchmarking workflow using **low-level TensorRT-LLM APIs** to characterize memory and performance bottlenecks, and to analyze precision–efficiency trade-offs across multiple quantization methods (RTN, AWQ, SmoothQuant) and runtime settings (batch size, sequence length, power modes).
- Analyzed FP16 inference failures (OOM at Batch=1) on 16GB Orin NX and benchmarked **W4A16 weight-only quantization** and **INT8 KV cache** to assess batch size and **sequence length scalability** under KV-dominated memory growth.
- Contributed to **TensorRT-LLM** (merged PR #9610 into `v0.12.0-jetson` branch) by adding a runtime guard clause to prevent memory profiler crashes on Jetson devices, ensuring stable benchmarking.

Projects

LLM Inference Optimization on T4 GPU | *GitHub*

- Engineered a high-efficiency inference pipeline for LLaMA-3.2-3B on a constrained T4 GPU environment, integrating Model Pruning, LoRA fine-tuning, Knowledge Distillation, and Mixed-Precision Quantization to minimize memory footprint.
- Accelerated end-to-end inference speed by over 5× while maintaining model quality (Perplexity ≤ 11.5).

Efficient Vision-Language Model via Multi-head Latent Attention | *GitHub*

- Implemented Multi-head Latent Attention (MLA) for LLaVA models by converting standard MHA to a low-rank latent structure with Partial RoPE and joint SVD-based KV compression.
- Achieved up to **80% KV cache memory reduction** with only **4% VQA accuracy drop**, enabling efficient inference on memory-constrained devices.

Skills

Languages: C/C++, Python, SQL, R

Tools: Git, Linux, Flex/Bison, Docker

ML & LLM Inference: PyTorch, TensorRT-LLM, vLLM, HuggingFace Transformers

Awards & Achievements

The 2023 ICPC Asia Taoyuan Regional Contest: Bronze Medal (Rank 41).

The 2023 ICPC Asia Taiwan Online Programming Contest: Silver Medal (Rank 38).

Extracurricular Activities

NYCU Programming Challenging Contest Association (PCCA) — *Hsinchu, Taiwan*
Member

- Engaged in weekly team training (3-5 hours) to solve complex algorithmic problems under time constraints.
- Solved over 1,500 problems on Codeforces and authored a comprehensive Team Codebook covering advanced algorithms and data structures for contest references.