Haechan An

+82-010-3943-3544 | haechan.an@kaist.ac.kr | anhaechan.github.io | haechan-an | AnHaechan

Introduction

Master's student at KAIST, graduating in Aug 2025. I work on **AI compilers** that will fill the gap of a software stack between large models, e.g., LLMs, and heterogeneous accelerators, e.g., NPU and PIM. I am trying to utilize my background on programming language design and theory for the good compiler design. I am seeking opportunities for solving industry-level problems for NPU compilers and overall AI systems.

Education

- **KAIST**, M.S. in School of Computing (expected) — Mar 2023 - Aug 2025 (expected)
  - Advisor: Jeehoon Kang
  - GPA: 3.87/4.3
- **KAIST**, B.S. in School of Computing — Mar 2018 - Feb 2023
  - GPA: 4.04/4.3 (Summa Cum Laude)
- **Yonsei University**, exchange studnet — Dec 2019 - Jan 2020

Skills

- **Programming languages:** Python, Rust, C++, Ocaml and Coq
- **AI systems:** PyTorch, TorchInductor, CUDA, LLVM and NeuPIMs simulator
- **Dev tools:** Linux CLI, Git, Docker and GCP

Projects & Experiences

- **AI compilers for NPU and PIM accelerators** — *Mar 2024 - Currently*
  - Wrote a survey on AI compilers for distributed and heterogeneous hardwares (submitted to a proprietary conference).
  - Created a lecture slide for introduction to AI compilers, contributing to KAIST CS492: Microarchitecture Design.
  - Wrote a code-level guide to TorchInductor, the backend AI compiler of PyTorch 2.
  - Implemented an extensible compiler framework, which takes PyTorch models then compiles them down to the binary, supporting simulation on NPU-PIM hardware.
  - Currently working on improving co-utilization of NPU-PIM hardwares during the LLM inference, by a novel data layout and scheduling strategy.
  - Currently working on designing an IR for NPU and GPU, which exposes hardware-level parallelism and data reuse.
  - Additionally, participated to SK Hynix's workshop on AiM (Accelerator-in-Memory) and Rebellions' PyTorch+NPU lab.
- **ML serving systems for NPU/GPU servers** — *Aug 2022 - Jan 2023*
  - Contributed to the evaluation of ML serving scheduler, by enabling GPU profiling and building an automatic compilation/profiling environment.
  - Been granted a patent on this topic, contributed as the third author.
- **Automatic program verification with language models and theorem provers** — *Mar 2023 - Feb 2024*
  - Studied and summarized the existing literature for proof automation.
  - Reproduced the work Thor (NeurIPS 2022), involving fine-tuning language models on TPU VMs and augmenting language model inference with a theorem prover.
- **Other projects**
  - Implemented a C static analyzer based on LLVM IR and the theory of abstract interpretation.
  - Implemented a C compiler that lowers C AST into IR, performs optimizations on IR, then generating an executable RISC-V assembly.
  - Implemented matrix multiplication in CUDA, using tiling and shared memory.

Honors and Awards

- **School of Computing, Dean's List:** Spring 2019, Fall 2020
- **School of Computing, Department Valedictorian:** Fall 2019
- **School of Computing, Excellent TA Award:** Spring 2023 (KAIST CS420 Compilers Design), Fall 2023 (KAIST CS220 Programming Principles)

Additional Information

- **Languages:** Korean (native), English (working proficiency), Japanese (intermediate)