인공지능 시스템 및 컴퓨터 구조 연구실

1. 지도교수: 오영환 (산학협력원 432호, 이메일: younghoh@ajou.ac.kr, 전화: 2368)

2. 연구분야

- System for Machine Learning (DNN Compiler, Library and Runtime)
- Model Serving System Optimization, Multi-tasking Scheduler for DNNs
- DNN Accelerator Architecture and Performance Modeling
- Hardware/Software Co-optimization

3. 학력

- 2013. 03 ~ 2022. 02 성균관대학교 전자전기컴퓨터공학과 박사 (컴퓨터 공학 전공)
- 2009. 03 ~ 2013. 02 성균관대학교 정보통신대학 전자전기공학과 학사

4. 주요경력

- 2022. 04 ~ 2024. 01 삼성종합기술원 Staff Researcher
  Universal Deep Learning Compiler (UDLC) Team, Computing Software T/U
- 2016. 04 ~ 2022. 02 서울대학교 컴퓨터공학과 방문연구원
  Architecture and Code Optimization (ARC) Lab

5. 논문

- 전체 논문 리스트는 홈페이지(https://aisa.ajou.ac.kr) 참조. 총 피인용 횟수 332회. (2024. 04 기준)

[IEICE TIS '22] "Layerweaver+: A QoS-aware Layer-wise DNN Scheduler for Multi-tenant Neural Processing Units", *IEICE Transactions on Information and Systems.*

[IEEE MICRO '21] "Accelerating Genomic Data Analytics with Composable Hardware Acceleration Framework", *IEEE MICRO: Special Issue on Top Picks from 2020 Computer Architecture Conferences.*

[HPCA '21] "Layerweaver: Maximizing Resource Utilization of Neural Processing Units via Layer-Wise Scheduling", *The 27th IEEE International Symposium on High Performance Computer Architecture.*

[ISCA '20] "Genesis: A Hardware Acceleration Framework for Genomic Data Analysis", *The 47th ACM/IEEE International Symposium on Computer Architecture* ***(IEEE Micro Top Picks 선정)***

[HPCA '20] "A3: Accelerating Neural Network Attention Mechanism with Approximation", *The 26th IEEE International Symposium on High Performance Computer Architecture.* ***(Google Scholar 기준 피인용 횟수 142+)***

[PACT '18] "A Portable Automatic Data Quantizer for Deep Neural Networks", *The 27th IEEE International Conference on Parallel Architectures and Compilation Techniques.* ***(Google Scholar 기준 피인용 횟수 36+)***

[ASPLOS '17] "Typed Architectures: Architectural Support for Lightweight Scripting", *The 22nd ACM Architectural Support for Programming Languages and Operating Systems* ***(ASPLOS Highlight Session 선정)***

[ISCA '16] "Short-Circuit Dispatch: Accelerating Virtual Machine Interpreters on Embedded Processors", *The 43rd IEEE/ACM International Symposium on Computer Architecture.*

[IEEE D&T '16] "An eDRAM-Based Approximate register File for GPUs", *IEEE Design & Test: Special Issues on Approximate Computing.*

[PPoPP-Poster '15] "JAWS: A JavaScript Framework for Adaptive CPU-GPU Work Sharing", *The 20th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming.*

[PRISM '14] "Automatic Runtime Selection of Best Hardware for Data-Parallel JavaScript Kernels via Lifelong Profiling", *The 2nd International Workshop on Parallelism in Mobile Platforms.*

6. 수상 내역

- 2022 Sukhan Lee Information-Intelligence Research Award (이석한 정보지능 연구 대상)
  성균관대학교 컴퓨터 관련 3개 분과에서 학위 기간 연구 업적으로 가장 우수한 박사 연구원으로 선정.
- 2021 IEEE MICRO Top Picks 2020
  2020년도 컴퓨터 구조 분야 최우수 학회(ISCA, HPCA, MICRO, ASPLOS 등)의 논문 중 Novelty 및 Long-term Impact를 평가하여 가장 우수한 12편의 논문 중 하나로 선정.
- ASPLOS Best Paper Nominee 2017
  컴퓨터 구조 분야 최우수 학회에서 53개 페이퍼 중 6개 최우수 논문 후보(Highlight Session)로 선정.

7. 연구실 (산학협력원 427호, 전화 : 2647, 홈페이지: https://aisa.ajou.ac.kr)

- 석사 2명
- 고성능 서버, GPU 및 개발 보드, 연구 환경(개인 PC, 듀얼 모니터, 모션 데스크, 연구 인센티브) 제공
- 석/박사 과정 학생 모집 – **"함께, 열정적으로, 즐겁게 논문 읽고 연구하실 분"**을 모집합니다.
  - 키트를 활용한 완성된 시스템 구현보다는 새로운 "시스템 최적화 기술 및 아이디어" 고안을 추구
  - **[Software]** Mobile/Embedded GPU 환경에서의 시스템 수준 최적화 (PyTorch 등의 내부 구현 수정)
  - **[Simulation]** 인공지능 모델의 효율적 실행을 위한 GPU 프로파일링 및 가벼운 성능 모델 개발
  - **[Hardware]** Verilog HDL, High-level Synthesis, Chisel 등을 이용한 FPGA 기반 가속 시스템 개발
  - **[Hardware]** CPU(RISC-V), GPU, NPU 등 구조에 대한 분석 및 이해를 통한 확장 모듈 설계
  - C++, Python, GPU Programming, Linux, DNN Framework/Compiler 관련 개발 경력 우대
    1) 효율적인 커뮤니케이션, 간결한 표현, 토론 능력
    2) 인공지능 시스템 관련 기술에 대한 높은 관심과 열정
    3) 영어 논문 읽기, 작성, 협업(기여 의지, 배려, 지식 습득 및 전파)

8. 연구 내용

1) Edge-to-Cloud Cooperative AI Serving System

- Speculative Decoding 등의 예측적 실행 기법을 활용한 Cloud-Edge 동시 실행 최적화
- AR/VR, Robot 등에서 활용되는 DNN 응용을 위한 멀티 태스킹 스케줄러 설계 및 성능 모델링
- Federated Learning 등 다수의 디바이스가 참여하는 On-device Inference/Training 최적화 연구로 확장

Keywords: Embedded GPU, Multi-tasking, Resource Scheduling, Parallel/Speculative Execution

2) Light-weight Performance Modeling for DNN Compilers

- GPU 등의 가속기 성능 프로파일링 및 코드 분석 기법을 활용하여 실험적/이론적 성능 분석
- 임베디드 환경에서 고정확성, 저전력으로 구동될 수 있는 런타임 성능 모델링 기법 개발
- 가벼운 성능 모델을 이용한 IoT향 인공지능 시스템에서의 실행/메모리 스케줄 최적화

Keywords: Analytical Modeling, Profiling, DNN Performance Estimation, DNN Compiler

3) Hardware & Software Co-optimized Accelerator Architecture

- 동적 구조의 DNN 모델(예, Language Model, Mixture-of-Experts 등)의 효율적 실행 기법 개발
- PyTorch 2.0 등 Just-in-Time DNN Compiler 동작 방식의 분석을 통한 하드웨어 가속 아이디어 고안
- FPGA 등을 통한 RISC-V 프로세서 확장 구조 설계 및 프로토타입 제작

Keywords: FPGA, JIT Compiler, Dynamic DNNs, Large Language Model, Gemmini (Berkeley)