라즈베리파이 5와 Hailo-8 및 8L의 AI 연산 성능 비교를 통한 엣지 디바이스 성능 향상 방안 연구

양병찬[1], 오세현[2], 차주형[2], 이제민[3], 권용인[3]
[1]경북대학교 전자공학부 학부생
[2]과학기술연합대학원대학교 인공지능학과
[3]한국전자통신연구원 온디바이스시스템SW연구실
ck27112@gmail.com,
(osehn, jh.cha, leejaymin, yongin.kwon)@etri.re.kr

A Study on Enhancing Edge Device Performance through AI Computation Comparison between Raspberry Pi 5 and Hailo-8 and 8L

Byeongchan Yang[1], Sehyeon Oh[2], Joohyoung Cha[2], Jemin Lee[3], Yongin Kwon[3]
[1]Dept. of Electronic Engineering, Kyungpook National University
[2]Dept. of Artificial Intelligence, University of Science and Technology
[3]On-Device System SW Laboratory, Electronics and Telecommunications Research Institute

요 약

본 연구는 엣지 컴퓨팅 환경에서 널리 사용되는 라즈베리파이 5의 AI 연산 성능 한계를 분석하고, 이를 보완할 수 있는 Hailo 기반 AI 가속기의 활용 가능성을 탐색하였다. 이미지를 분류를 위한 트랜스포머 기반 AI 모델 활용하여, 라즈베리파이 5를 사용한 경우와, Hailo-8 또는 Hailo-8L 모듈을 함께 사용한 경우의 연산 성능을 비교 분석하였다. 실험 결과, Vit Base BN 모델을 기준으로 Hailo-8을 활용했을 때 약 81.8%의 성능 향상이 나타났으며, 이를 통해 엣지 디바이스에서의 실시간 AI 처리 효율이 개선됨을 확인할 수 있었다.

1. 서론

최근 엣지 컴퓨팅 기술의 발전과 AI 응용 분야의 확산으로, 로봇, 드론, 스마트 카메라, 웨어러블 기기 등 다양한 모바일 디바이스에서 실시간 인공지능 추론을 수행하려는 수요가 지속적으로 증가하고 있다[1]. 특히 영상 인식, 객체 검출, 자연어 처리 등 복잡한 딥러닝 모델을 저지연으로 처리하기 위해, 연산 자원이 제한된 디바이스에서도 고성능 AI 처리를 가능하게 하는 기술의 필요성이 대두되고 있다[2]. 라즈베리파이는 저전력, 소형, 저비용의 장점을 바탕으로 다양한 임베디드 시스템에서 활용되고 있으며, 최근 출시된 라즈베리파이 5는 Arm Cortex-A76 기반 CPU와 PCIe 확장 인터페이스를 제공함으로써 엣지 AI 애플리케이션의 구현 가능성을 높이고 있다. 그러나 라즈베리파이 5 단독으로는 트랜스포머 기반의 복잡한 딥러닝 모델을 실시간으로 추론하기에는 여전히 한계가 존재한다.

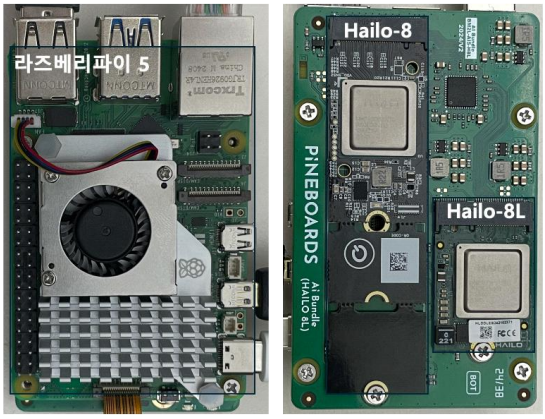
이에 본 연구에서는 고성능 엣지 AI 가속기인 Hailo-8과 경량 버전인 Hailo-8L을 라즈베리파이 5에 연결하여, 트랜스포머 기반 이미지 분류 모델에 대한 추론 성능을 비교 및 분석하였다. 실험에는 Hailo Model Zoo에서 제공하는 사전 학습된 모델을 사용하였다. 라즈베리파이 5에 Hailo-8과 Hailo-8L 모듈을 함께 연결한 단일 하드웨어 환경에서, 라즈베리파이 5 단독 실행, Hailo-8 활용, Hailo-8L 활용의 세 가지 방식으로 추론 성능을 측정하였다. 이를 통해 외장형 AI 가속기가 엣지 디바이스 성능에 미치는 영향을 정량적으로 평가하고자 한다.

2. 시스템 구성 및 실험환경

표 1은 본 연구에 사용된 하드웨어 및 소프트웨어 환경을 나타낸다. 제어 장치로는 라즈베리파이 5를 사용하였으며, 이는 Arm Cortex-A76 쿼드코어 프로세서와 8GB LPDDR4X RAM을 탑재하고 있다. 라즈베리파이 5는 PCIe 2.0 x1 레인 인터페이스를 통해 외부 장치와의 고속 통신이 가능하다.

본 연구에서는 그림 1과 같이 M.2 HAT+ 보드를 활용하여 PCIe 인터페이스를 통해 Hailo-8과 Hailo-

8L 모듈과 연결하였다. 이를 통해 단일 시스템에서 두 종류의 Hailo 모듈을 모두 사용할 수 있는 환경을 구성하였다.



(그림 1) 실험 환경.

Hailo-8은 최대 26 TOPS의 AI 추론 성능과 2.5 W의 저전력 소모를 제공하는 엣지 AI 프로세서로, CNN에 최적화되고 트랜스포머 기반의 모델도 지원한다. Hailo-8L은 13 TOPS의 성능과 1.5W로 가벼운 애플리케이션에 적합한 저전력 프로세서로, 두 프로세서 모두 Hailo Dataflow Compiler와 HailoRT를 활용해 개발 효율성을 높일 수 있다.

<표 1> 하드웨어 및 소프트웨어 사양

구분	항목	라즈베리파이 5	Hailo-8	Hailo-8L
H/W	프로세서 구성	2.4GHz ARM Cortex-A76 MP4 CPU	AI 프로세서	AI 프로세서
	최대 연산 성능	31.4 GFLOPS	26 TOPS	13 TOPS
	전력 소비량	2.6-7W	2.5W	1.5W
S/W	운영체제	Raspberry Pi OS	Linux 기반 호스트 시스템 제어	Linux 기반 호스트 시스템 제어
	프레임워크	onnxruntime	HailoRT	HailoRT

본 실험은 총 세 가지 환경을 기반으로 수행되었다. 첫 번째는 라즈베리파이 5 단독으로 AI 모델을 실행한 경우며, 두 번째와 세 번째는 라즈베리파이 5에서 각각 Hailo-8과 Hailo-8L 모듈을 활용하여 모델을 실행한 경우이다. 각 방식에서는 이미지 분류를 위한 트랜스포머 기반 AI 모델을 사용하였다. 라즈베리파이 5 단독 실행에 대한 추론 성능은 onnxruntime을 이용하여 측정하였으며, Hailo-8 및 Hailo-8L을 사용한 경우에는 HailoRT API를 활용하였다. 이때 정확한 성능 측정을 위해 1회 워밍업 후 20회 반복 실행하여 평균을 계산하였다.

3. 실험 결과

<표 3> AI 모델의 추론 시간 성능 비교

모델	라즈베리파이 5(ms)	Hailo-8(ms)	Hailo-8L (ms)
DeiT Tiny	55.1	26.5 (-51.9%)	31.8 (-42.3%)
ViT Base BN	838.0	152.2 (-81.8%)	157.0 (-81.3%)
ViT Small	225.8	59.94 (-73.5%)	63.6 (-71.8%)
ViT Tiny	86.8	24.4 (-71.8%)	31.7 (-63.5%)

실험 결과, 트랜스포머 기반 모델에서 Hailo를 활용한 추론 환경이 라즈베리파이 5 단독 실행 대비 높은 성능 향상을 보였다. 특히 Vit Base BN의 경우 Hailo-8을 사용할 때 추론 시간이 81.8%의 성능 개선을 기록하였다. 이는 트랜스포머와 같이 구조가 복잡한 모델에 대해서도 Hailo 기반 AI 가속기가 효과적으로 동작함을 보여준다.

4. 결론

본 연구에서는 라즈베리파이 5에 외장형 AI 가속기를 연동하여 엣지 디바이스 기반 인공지능 연산 환경을 구현하였다. 실험 결과 Hailo-8과 같은 외장형 AI 모듈이 경량 디바이스 기반 AI 시스템 성능을 실질적으로 개선할 수 있음을 실험적으로 입증하였다. 향후에는 해당 시스템을 자율주행 등 실시간성이 요구되는 응용 분야에 적용하여 온디바이스 환경에서의 활용 가능성을 검토할 예정이다.

감사의 글

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구 결과 임 (No.RS-2024-00459797, No.RS-2023-00277060)

참고문헌

[1] Singh Raghubir, Gill Sukhpal Singh, Edge AI: a survey, Internet of Things and Cyber-Physical Systems, 3, 71 - 92, 2023.

[2] Véstias Mário P., Duarte R. P., de Sousa, J. T., Neto, H. C., Moving deep learning to the edge, Algorithms, 13, 5, 125, 2020