엣지 컴퓨팅용 임베디드 보드의 PyTorch 기반 객체인식 모델 추론 성능 비교

유채원, 오명훈
호남대학교 컴퓨터공학과
kelly5233@naver.com, mhoh@honam.ac.kr

Inference Performance Comparison of PyTorch-based Object Detection Models on Embedded Boards for Edge Computing

Chae-Won You, Myeong-Hoon Oh
Dept. of Computer Engineering, Honam University

요 약

본 논문은 PyTorch 기반 YOLOv8 모델을 Raspberry Pi 5, LattePanda 3 Delta, Jetson Nano 등 다양한 엣지 컴퓨팅용 임베디드 보드에 적용하여 추론 성능을 비교하였다. 실험 결과, PyTorch의 Intel CPU 최적화 영향으로 LattePanda 3 Delta가 가장 우수한 속도를 기록하였으며, 이는 하드웨어뿐 아니라 딥러닝 프레임워크 최적화 여부 또한 성능에 중요한 영향을 미친다는 점을 보여준다.

Ⅰ. 서 론

최근 인공지능 기술의 발전과 함께 이미지 인식 기술은 다양한 산업 분야에서 핵심적인 역할을 수행하고 있다. 위성사진 분석을 통한 원유 저장 탱크의 산출량 추정, 사용자 이미지 기반의 유사 제품 검색 등 다양한 응용 사례가 보고되고 있다. 이러한 기술들은 딥러닝과 신경망 알고리즘을 기반으로 하여 사람의 나이, 성별, 기분까지도 인식할 수 있는 수준에 이르렀다. 특히, 이미지 인식 기술은 산업 전반에 걸쳐 적용되고 있다.[1]

이미지 인식 속에서도 객체탐지는 컴퓨터 자원의 발전과 더불어 지속적으로 대두되었다. 그중에서도 YOLO(You Only Look Once)는 다양한 버전을 통해 객체탐지 모델의 주류가 되어갔다[1].

기존의 이미지 인식 및 객체 탐지 기술의 학습을 위해서는 주로 고성능 서버나 GPU 환경에서 운용되어 왔다. 이러한 장비들은 알고리즘이 요구하는 연산량을 충족시키기 위해 고사양 하드웨어를 필요로 한다. 그러나, 엣지 컴퓨팅용 소형 임베디드 보드과 같은 제한된 자원 환경에서는 적용에 제약이 따른다[2]. 반면,YOLOv8이 Raspberry Pi 5와 같은 소형 보드에서도 추론이 가능함을 보였다[3].

본 논문에서는 LattePanda 3 Delta, Raspberry Pi 5, Jetson Nano와 같은 다양한 임베디드 보드에서 대표적인 객체 인식 프레임워크인 PyTorch[4] 기반 모델을 적용하여, 실시간 추론 성능을 비교·분석하고자 한다.

Ⅱ. 본 론

본 논문에서는 YOLOv8 계열 중 가장 경량화된 모델인 YOLOv8n(nano)을 사용하여 객체 탐지 성능을 비교하였다. YOLOv8n은 Ultralytics에서 제공하는 YOLOv8 시리즈의 경량화 버전으로, 연산량이 적고 모델 크기가 작아 임베디드 보드에 적합한 추론 효율을 제공한다[5][6]. 실험은 Python 3.8.13, PyTorch 2.4.1 환경에서 수행되었으며, 입력 이미지의 크기는 640x640으로 설정하였다. 이는 기존 연구들에서 기본값 혹은 실험용 기준값으로 자주 사용되는 해상도이다[1].

탐지 후처리 과정에서는 IoU(Intersection over Union) 임계값을 0.45로, confidence threshold는 0.15로 설정하였다. 또한 실시간 탐지 결과를 확인하기 위해 시각화 옵션인 show=True를 활성화하여, 각 장치에서의 추론 결과를 화면에 출력하였다.

추론 테스트용 데이터셋으로는 COCO128을 사용하였다. COCO128은 COCO 데이터셋에서 일부 이미지를 추출하여 구성한 경량 테스트용 샘플 데이터셋으로, Ultralytics에서 공식 배포하고 있으며, 모델 크기와 연산량이 작아 임베디드 보드 환경에 적합하다는 점에서 본 연구의 비교 실험에 사용되었다[7].

실험에 사용된 코드는 GitHub에 공개된 YOLOv8 기반 개인 리포지터리인 jetsonmom/yolov8_jetson4GB를 참고하여 구성되었으며, 전체 구조와 설정은 대부분 유지하되, 데이터 경로 등 일부 항목만 실험 목적에 맞게 간단히 수정하여 사용하였다[8].

각 장치에서의 실시간 추론 성능은 YOLOv8 추론 코드를 30초 동안 실행하여 해당 기간 동안의 처리 속도의 평균값과 중앙값을 측정하여 평가하였다.

그림 1. Raspberry Pi 5에서 YOLOv8 추론 실행 결과 화면

그림 2. Jetson Nano에서 YOLOv8 추론 실행 결과 화면

그림 3. LattePanda 3 Delta에서 YOLOv8 추론을 실행한 결과 화면

그림 1, 2, 3은 각각 LattePanda 3 Delta, Raspberry Pi 5, Jetson Nano에서 YOLOv8n 모델을 이용해 실시간 객체 탐지를 수행한 결과를 나타낸다. 각 장치에서 출력된 bounding box와 confidence score를 통해 모델이 정상적으로 작동함을 확인할 수 있었다. 세 장치 간 탐지 결과는 시각적으로 유사한 양상을 보였으며, 주어진 조건에서 실험이 안정적으로 잘 수행되었음을 보여준다.

표 1. YOLOv8 추론 성능 비교 (30초 기준)

장치	평균 처리 시간(ms)	중앙값 처리 시간(ms)
LattePanda 3 Delta	217.75	217.75
Raspberry Pi 5	493.71	487.5
Jetson Nano	704.44	665.05

표 1은 YOLOv8n 추론 코드를 30초 동안 실행하여 각 임베디드 보드에서 측정한 평균 및 중앙값 처리 시간을 비교한 것이다. 수치적으로도 LattePanda 3 Delta, Raspberry Pi 5, Jetson Nano 순으로 추론 속도가 빨랐다. LattePanda 3 Delta는 Jetson Nano보다 약 69%, Raspberry Pi 5 보다 약 56% 더 빠른 속도를 보였다.

표 2. 실험에 사용된 임베디드 보드의 CPU 및 메모리사양[9-11]

장치	CPU	메모리
LattePanda 3 Delta	Intel Celeron N5105, Quad-core, 2.0 - 2.9GHz	8GB
Raspberry Pi 5	Broadcom BCM2712, Cortex-A76, Quad-core 2.4GHz	8GB
Jetson Nano	ARM Cortex-A57, Quad-core, 1.43GHz	4GB

표 2는 실험에 사용된 임베디드 보드에 탑재된 CPU 사양 및 메모리를 정리한 것이다. 세 보드 중 LattePanda 3 Delta와 Raspberry Pi 5는 8GB 메모리를 탑재하고 있으며, Jetson Nano는 기본 모델 기준 4GB 메모리를 사용하였다. Jetson Nano는 상대적으로 적은 메모리 용량과 낮은 CPU 성능으로 인해, 실험에서 가장 낮은 추론 성능을 기록하였다. 이 결과는 메모리 용량 역시 성능에 직접적인 영향을 미칠 수 있음을 보여준다.

Raspberry Pi 5는 고클럭(2.4GHz)의 Cortex-A76) 기반 프로세서를 탑재하고 있음에도 불구하고, 실제 처리 속도에서는 Intel 아키텍처 기반의 LattePanda 3 Delta가 가장 우수한 성능을 보였다. 이는 PyTorch 프레임워크가 Intel 기반 시스템에 보다 최적화되어 있기 때문으로 해석되며, 따라서 본 실험에서의 성능 차이는 메모리보다는 CPU 아키텍처와 프레임워크의 연산 최적화 정도에 기인한 것으로 판단된다.

Ⅲ. 결 론

본 논문에서는 YOLOv8 객체 탐지 모델을 PyTorch 기반으로 각기 다른 임베디드 보드에서의 추론 성능을 비교하였다. 실험 결과, PyTorch가 Intel 아키텍처에 최적화되어 있다는 점에서 LattePanda 3 Delta가 가장 높은 추론 속도를 기록하였다.

평균 처리 시간 기준, LattePanda 3 Delta는 217.75ms, Raspberry Pi는 493.71ms, Jetson Nano는 704.44ms를 각각 기록하였으며, LattePanda 3 Delta는 Jetson Nano보다 약 69%, Raspberry Pi 5 보다 약 56% 더 빠른 속도를 보였다. 이는 임베디드 환경에서 AI 모델 적용 시, 하드웨어 스펙뿐만 아니라 소프트웨어 프레임워크의 최적화 여부가 중요한 성능 결정 요소가 될 수 있음을 시사한다.

다만 본 실험은 PyTorch의 CPU 전용 빌드 환경에서 수행되었으며, 향후에는 GPU 기반 CUDA 환경 및 다양한 모델에 대한 후속 실험이 필요하다.

ACKNOWLEDGEMENT

이 논문은 2025년도 과학기술정보통신부의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임. (No.2022-0-010000, 이동형 맞춤 의료서비스 지원을 위한 유연의료 5G 엣지 컴퓨팅 SW 개발)

참 고 문 헌

[1] Terven, J., and Cordova-Esparza, D,. “A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS,” arXiv preprint, arXiv:2304.00501, Apr. 2023

[2] Capra, M., Bussolino, B., Marchisio, A., Masera, G., Martina, M., & Shafique, M., “Hardware and Software Optimizations for Accelerating Deep Neural Networks: Survey of Current Trends, Challenges, and the Road Ahead,” arXiv preprint, arXiv:2012.11233, Dec. 2020

[3] Alqahtani, D. K., Cheema, A., and Toosi, A. N., “Benchmarking Deep Learning Models for Object Detection on Edge Computing Devices,” arXiv preprint arXiv:2409.16808, Sep. 2024

[4] Intel, “PyTorch* Optimizations from Intel,” Intel Developer, https://www.intel.com/content/www/us/en/developer/tools/oneapi/optimization-for-pytorch.htmlaccessed, May. 2, 2025

[5] Ultralytics, “Explore Ultralytics YOLOv8,” Ultralytics YOLO Docs, https://docs.ultralytics.com/ko/models/yolov8/, accessed Apr. 30, 2025

[6] Sapkota, R., & Karkee, M., “Comparing YOLOv11 and YOLOv8 for Instance Segmentation of Occluded and Non-Occluded Immature Green Fruits in Complex Orchard Environment,” arXiv preprint, arXiv:2410.19869, Jan. 2025

[7] Ultralytics, “COCO128 Dataset” Ultralytics YOLO Docs, https://docs.ultralytics.com/ko/datasets/detect/coco128/, accessed Apr. 30, 2025

[8] jetsonmom, “yolov8_jetson4GB,” GitHub repository, https://github.com/jetsonmom/yolov8_jetson4GB, accessed Apr. 30, 2025

[9] LattePanda, “LattePanda 3 Delta,” LattePanda Official Webside, https://www.lattepanda.com/lattepanda-3-delta, accessed May. 2, 2025

[10] Raspberry Pi , “Raspberry Pi 5,” Raspberry Pi Official Webside, https://www.raspberrypi.com/products/raspberry-pi-5/, accessed May. 2, 2025

[11] NVIDIA, “Jetson Nano” NVIDA Developer, https://developer.nvidia.com/embedded/jetson-nano, accessed May. 2, 2025