투명 페트병 분리를 위한 Hailo-8 기반 경량 객체 검출 시스템

윤태준[1O], 정윤희[1], 이소연[2], 김대영[1*]
[1]순천향대학교 컴퓨터소프트웨어공학과
[2]순천향대학교 소프트웨어융합학과
e-mail : {*20214004, yh9652, lsy8647, dyoung.kim*}*@sch.ac.kr*

A Hailo-8-Based Lightweight Objectct Detection System for Transparent PET Bottle Separation

Tae Jun Yoon[1O], YunHee Jeong[1], SoYeon Lee[2], Dae-Young Kim[1*]
[1]Department of Computer Software Engineering, Soonchunhyang University
[2]Department of Software Convergence, Soonchunhyang University

Abstract

The increasing use of single-use plastics has raised the need for automated PET bottle separation systems to ensure high-quality recycling. This study proposes a lightweight AI-based system that detects key components of transparent PET bottles—such as caps, rings, and labels—in real time to improve separation accuracy. A previous Jetson Nano-based system exhibited limitations in real-time performance, achieving an average of only 3.16 FPS. To address this, we integrated the Hailo-8 AI inference accelerator with a Raspberry Pi 5 and applied 8-bit quantization techniques after converting a YOLOv8n model into ONNX format. The proposed system achieved an average inference speed of 27.18 FPS, representing an 8.6× improvement over the baseline. The results demonstrate that the system operates reliably even in embedded environments, and it holds potential for extension into parallel inference and multi-object detection applications in real-world recycling scenarios.

I. 서론

전 세계적으로 일회용 플라스틱 사용량이 급증함에 따라, 플라스틱 폐기물은 심각한 환경 문제로 부상하고 있다. 특히 투명 페트병은 가장 많이 사용되는 일회용 플라스틱 중 하나로, 효율적인 재활용을 위해서는 뚜껑과 라벨의 정확한 분리가 필수적이다. 한국소비자원에 따르면, 고부가가치 재활용이 가능한 부위는 주로 페트병 몸통이며, 이를 위해 자동화된 분리 시스템의 도입이 요구된다 [1].

그러나 기존 분리 공정은 대부분 수작업에 의존하고 있어, 분리 정확도 저하, 인건비 상승, 대량 처리의 어려움 등 여러 한계가 존재한다. 이를 해결하기 위해 본 연구팀은 Jetson Nano 기반의 딥러닝 객체 검출 기술을 활용하여, 투명 PET 병의 구성 요소를 실시간으로 검출 및 분류하는 시스템을 개발하였다 [2]. 해당 시스템은 기본적인 실시간 추론은 가능했으나, 평균 3.16FPS 추론 속도로 인해 고속 처리가 요구되는 응용 환경에 적용하기에는 한계가 있었으며, 연산 효율성과 처리 속도 측면에서의 개선이 요구되었다.

이에 본 연구에서는 고성능 AI 추론을 위한 경량화 칩셋인 Hailo-8을 활용하여 기존 시스템의 실시간성 및 연산 효율성을 향상시키고자 한다. Hailo-8은 높은 연산 효율성과 실시간 추론 성능을 동시에 만족시킬 수 있는 구조를 기반으로 하며, 이를 Raspberry Pi 5와 연계하여 소형 저비용의 고성능 투명 PET 분류 시스템을 구현하였다. 본 논문에서는 해당 시스템의 설계 및 구현 과정을 설명하고, 기존 Jetson Nano 기반 시스템 대비 추론 속도 및 임베디드 환경에서의 적용 가능성 측면에서 성능 개선 결과를 실험적으로 분석한다.

II. 관련 연구

2.1 Hailo-8

최근 고속 실시간 추론에 대한 수요가 증가함에 따라, 엣지 디바이스 환경에서도 높은 연산 성능과 에너지 효율을 동시에 만족하는 AI 가속기의 필요성이 증가하고 있다. Hailo-8은 이러한 요구를 충족하기 위해 설계된 경량형 고성능 AI 추론 칩셋으로, 최대 26 Tera Operations Per Second(TOPS)의 연산 성능과 와트 당 2.8TOPS의 전력 효율을 제공한다 [3]. 기존 CPU 및 GPU 기반 구조와 달리, Hailo-8은 외부 메모리 의존을 제거하고 메모리, 연산 유닛, 제어 유닛을 칩 내부에 분산 배치한 구조를 채택하여 데이터 이동에 따른 오버헤드를 최소화하였다. 이러한 구조는 연산 지연을 줄이고, 전력 소비를 절감하며, 복합 객체 검출 모델의 실시간 처리를 가능하게 하여 소형 임베디드 시스템에 적합한 추론 플랫폼으로 주목받고 있다.

2.2 양자화 기법과 대표값 보정

Hailo-8 기반 추론 시스템을 구현하기 위해서는, 사전 학습된 AI 모델을 정수 형태로 변환하는 양자화 과정이 필수적으로 수행된다. 양자화는 모델의 가중치 및 활성화 값을 8bit 정수값으로 압축함으로써, 전체 모델 크기를 축소하고 연산 효율을 향상시킬 수 있는 대표적인 경량화 기법이다. 특히, 양자화 성능의 핵심은 각 레이어의 활성화 범위 추정(calibration)에 있으며, 이는 양자화 오차를 최소화하는 데 중요한 역할을 한다 [4].

본 연구에서는 Hailo-8의 Random Calibration 옵션을 사용하여, 실제 입력 데이터셋 없이 무작위로 생성된 RGB 입력을 기반으로 각 레이어의 활성화 범위를 추정하였다. 해당 방식은 별도의 대표 입력 데이터셋 없이도 양자화를 수행할 수 있으며, 개발 초기 단계에서 모델 경량화 및 추론 속도 검증에 유리한 접근 방식이다. 다만, 실제 데이터 분포를 반영하지 않기 때문에 양자화 정확도에는 제한이 있을 수 있다 [5].

2.3 ONNX 변환과 Hailo Dataflow Compiler

PyTorch 또는 TensorFlow 등 다양한 프레임워크에서 개발된 모델을 Hailo-8에서 실행하기 위해서는, 먼저 모델을 Open Neural Network Exchange(ONNX) 형식으로 변환해야 한다. ONNX는 다양한 프레임워크 간의 상호 운용성을 제공하며, 내부 연산 그래프 최적화를 통해 추론 속도 향상에도 기여하는 범용 표준 포맷이다 [6]. 변환된 ONNX 모델은 Hailo에서 제공하는 Hailo Dataflow Compiler를 통해 Hailo 고유 실행 포맷인 Hailo Executable Format(HEF)으로 컴파일 되며, 이는 Hailo 장치에서 직접 실행 가능한 바이너리 파일이다. 본 연구에서는 해당 과정을 통해 YOLOv8n 기반 객체 검출 모델을 HEF 형식으로 컴파일했으며, Raspberry Pi 5와 Hailo-8을 연동한 임베디드 환경에서 실시간 추론 성능을 실험적으로 검증하였다.

III. 시스템 설계 및 학습 결과

3.1 시스템 구성

본 연구에서는 PET 병의 구성 요소를 실시간으로 검출하고 시각적으로 안내함으로써 재활용 품질을 향상시키기 위한 경량 AI 기반 객체 검출 시스템을 설계하였다. 전체 시스템의 구조는 그림 1에 나타나 있으며, 시스템은 (1) 객체 검출 모델을 학습하는 AI Server, (2) 학습된 모델을 양자화하고 Hailo에 최적화된 실행 포맷으로 변환하는 Quantization Server, (3) 실시간 추론을 수행하는 Raspberry Pi 5 + Hailo-8 추론 장치로 구성된다.

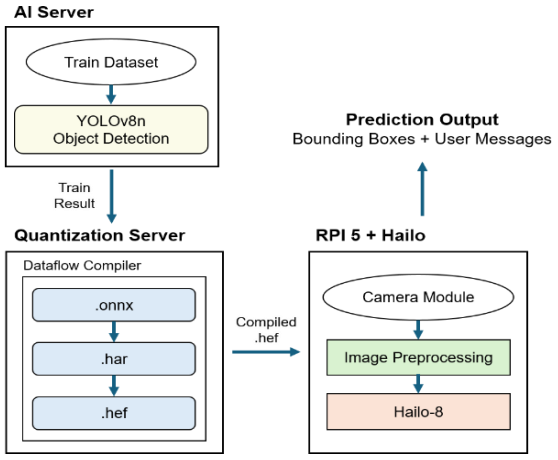

그림 1. 시스템 구조

전체 시스템의 데이터 흐름은 그림 2에 요약되어 있으며, 모델 학습부터 양자화 및 추론, 사용자 안내 메시지 출력까지의 과정이 정리되어 있다.

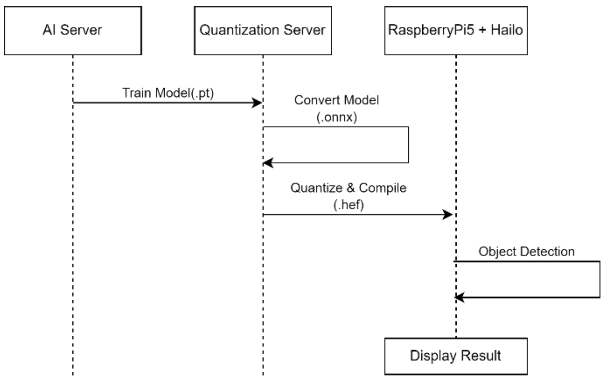

그림 2. 시스템 흐름도

AI Server에서는 YOLOv8n 기반의 객체 검출 모델을 학습하며, PET 병의 구성 요소인 뚜껑(cap), 고리(ring), 라벨(label)을 대상으로 한다. 학습에는 AI Hub에서 제공하는 생활 폐기물 이미지 데이터셋을 활용하였으며, 구성 요소별로 Bounding Box 어노테이션을 적용하였다. 학습이 완료된 모델은 ONNX 형식으로 변환된 후 Quantization Server로 전달되며, Hailo Dataflow Compiler를 통해 .hef 실행 포맷으로 양자화 및 컴파일 된다. 이 과정에는 Random Calibration 방식의 8bit 정수 양자화가 적용되어, Hailo-8 장치에서 고속 추론이 가능한 경량 모델이 생성된다.

양자화가 완료된 모델은 Raspberry Pi 5 장치로 전송되며, 실시간으로 입력되는 카메라 영상을 기반으로 객체 검출을 수행한다. 검출 결과는 Bounding Box와 함께 시각적으로 출력되며, 사용자에게는 해당 구성 요소의 제거를 유도하는 안내 메시지가 제공된다.

3.2 객체 검출 및 모델 학습 결과

본 연구에서 활용된 객체 검출 모델은 기존 연구 [2]에서 학습된 YOLOv8n 기반의 모델을 기반으로 하며, PET 병의 주요 구성 요소인 뚜껑(cap), 고리(ring), 라벨(label)을 검출할 수 있도록 설계되었다. 학습에는 AI Hub 생활 폐기물 이미지 데이터셋이 활용되었으며, CVAT를 활용한 Bounding Box 어노테이션이 적용되었다. 모델 학습 결과는 기존 연구에 상세히 보고되어 있으며, Can/Plastic 및 Cap/Ring/Label 객체에 대해 각각 0.99 및 0.88의 mAP@0.5를 기록하였다.

본 논문에서는 해당 모델을 실시간 임베디드 환경에 적용하기 위해, 기존 모델을 ONNX 형식으로 변환한 후 Hailo Dataflow Compiler를 통해 .hef 실행 포맷으로 양자화 및 최적화하였다. 특히, Random Calibration 기반의 8bit 정수 양자화를 적용함으로써 Hailo-8 장치에서의 고속 경량 추론을 가능하게 하였다.

Ⅳ. 실험 결과

본 연구에서는 제안한 PET 병 구성 요소 객체 검출 시스템의 실시간성 성능을 평가하기 위해, 기존 Jetson Nano 기반 시스템과 Hailo-8 기반 시스템 간의 평균 프레임 처리 속도(Frames Per Second, FPS)를 비교하였다. 실험은 동일한 학습 모델과 입력 영상 조건에서 수행되었으며, 두 시스템의 평균 추론 속도를 측정하여 성능 차이를 정량적으로 분석하였다.

Jetson Nano 기반 시스템은 평균 3.16 FPS를 기록하였으며, 이는 실시간 처리가 요구되는 응용 환경에서는 한계가 있는 수준이다. 반면, Raspberry Pi 5와 Hailo-8을 기반으로 구성된 개선된 시스템은 그림 3과 같이 평균 27.18 FPS를 기록하였고, 기존 대비 약 8.6배 향상된 추론 속도를 보였다. 이를 통해 제안 시스템은 PET병의 구성 요소 검출을 실시간으로 수행하면서도, 임베디드 환경에서 안정적인 처리 성능을 확보할 수 있음을 입증하였다.

표 1은 두 시스템의 평균 FPS를 정량적으로 비교한 결과로, 제안 시스템이 Jetson Nano 대비 실시간 추론 성능에서 개선 효과를 보였음을 확인할 수 있다.

그림 3. 구현 결과

표 1. 실험 결과 분석

시스템 구성	평균 FPS
Jetson Nano	3.16
Raspberry Pi 5 + Hailo-8	27.18

Ⅳ. 결론 및 향후 연구

본 연구에서는 기존 Jetson Nano 기반 PET 분류 시스템의 실시간성 한계를 극복하기 위해, Hailo-8 칩셋을 활용한 추론 구조를 설계하고, 학습된

YOLOv8n 모델을 ONNX 변환 및 양자화 과정을 통해 임베디드 환경에 최적화하였다.

실험 결과, 제안한 Raspberry Pi 5 + Hailo-8 기반 시스템은 평균 27.18 FPS의 추론 성능을 기록하였으며, 이는 기존 시스템 대비 약 8.6배 향상된 속도로, 실시간 검출 및 안내 메시지 출력이 가능한 수준의 처리 성능을 확보하였음을 보여준다. 또한, 제안 시스템은 경량화된 임베디드 환경에서도 안정적으로 동작함을 실험적으로 확인하였다.

향후 연구에서는 복합 객체 검출 모델의 실시간 추론 성능을 더욱 향상시키기 위해, 병렬 추론 구조를 설계하고 이를 기반으로 복합 객체에 대한 동시 분석 및 출력이 가능한 시스템으로 확장할 예정이다.

* 본 연구는 2025년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음 (2021-0-01399)

참고문헌

[1] 한국소비자원, "투명(무색)페트병 분리배출 소비자문제 조사," Korean Consumer Agency, 2021.

[2] 정윤희, 이소연, 김대영, "재활용 비용 절감을 위한 실시간 투명 페트병 검출 및 분류 시스템," 한국통신학회 학술대회 논문집, pp. 1382-1384, 2025.

[3] 허두환, 박대현, 김덕웅, 배재용, 박준형, 배승환, "군용 도메인 영상에 대한 서버와 온-보드 간의 객체 검출 성능 분석," 한국컴퓨터정보학회 논문지, vol. 29, no. 8, pp. 157-164, 2024.

[4] 권도혁, 유연호, 조휘주, 최현목, 유혁, 양경식, "네트워크 트래픽 사이즈 예측 모델의 경량화를 위한 양자화 기법 분석 및 비교," 한국정보과학회 학술발표논문집, pp. 1677-1679, 2024.

[5] Hailo Technologies Ltd., "Hailo Dataflow Compiler User Guide," Version 3.27.0, March 2024.

[6] 최다훈, 이종호, 홍현석, 강희범, 김현, "ONNX 변환을 통한 심층신경망의 성능 분석," 대한전자공학회 하계학술대회, pp. 1085-1086, 2023.