Transactions of KSAE, Vol. 34, No. 9, pp.1053-1061 (September, 2026)

pISSN 1225-6382 / eISSN 2234-0149
DOI http://dx.doi.org/10.7467/KSAE.2026.34.9.1053

〈응 용 논 문〉

임베디드 AI 객체 탐지 모델의 전력 효율 분석 및 평가 프레임워크 연구

유진호[1] · 기석간[2] · 기석철*[3]

에이브이지니어스 연구개발팀[1] · 건국대학교 메카트로닉스공학과[2] · 충북대학교 지능로봇공학과[3]

A Power-Efficiency Analysis and Evaluation Framework for Embedded AI-Based Object Detection Models

Jinho Yoo[1] · Seokkan Ki[2] · Seok-Cheol Kee*[3]

[1]R&D Team, AVGenius Co., 45 Yangcheong 4-gil, Ochang-eup, Cheongwon-gu, Cheongju-si, Chungbuk 28116, Korea
[2]Department of Mechatronics Engineering, Konkuk University, Chungbuk 27478, Korea
[3]Department of Intelligent Systems and Robotics, Chungbuk National University, Chungbuk 28644, Korea
(Received 27 February 2026 / Revised 27 March 2026 / Accepted 27 March 2026)

Abstract : In automotive applications, embedded artificial intelligence (AI) systems must maintain real-time perception while operating within strict power, thermal, and computational constraints. However, most existing studies on object detection, particularly those using YOLO architectures, emphasize detection accuracy (mAP) and inference speed (FPS), yet give far less attention to energy efficiency, a critical factor for deployment on vehicle-mounted embedded platforms.

This study proposes a power-efficiency-oriented evaluation framework that jointly considers detection accuracy, inference throughput, power consumption, and energy efficiency expressed as mAP per Watt to overcome this limitation. YOLOv5 serves as the representative object detection model, and two widely adopted model compression techniques, structured pruning and post-training quantization, are systematically applied and evaluated. Using a vehicle-mounted embedded platform, experiments are carried out under the same operating conditions. The results indicate that models with similar detection accuracy may display markedly different power consumption characteristics, while quantized models deliver the highest energy efficiency with only minimal accuracy degradation. These findings underscore the importance of energy-aware evaluation for practical automotive embedded perception systems.

Key words : Embedded AI(임베디드 인공지능), Object detection(객체 탐지), Pruning(가지치기), Quantization(양자화), Power efficiency(전력 효율), mAP/W(에너지 효율 지표)

1. 서 론

차량 기술의 고도화와 함께 첨단 운전자 지원 시스템(Advanced Driver Assistance Systems, ADAS) 및 자율주행 시스템에서 주변 환경을 인식하기 위한 객체 탐지 기술의 중요성이 지속적으로 증가하고 있다.[1,2] 카메라 기반 객체 탐지는 차량 전방 및 주변의 차량, 보행자, 장애물 등을 인식하는 핵심 요소로 활용되며, 최근에는 딥러닝 기반 객체 탐지 모델이 널리 적용되고 있다. 특히 YOLO 계열 객체 탐지 모델은 단일 단계 검출 구조를 기반으로 높은 실시간성을 제공하여 차량 적용 가능성이 높은 것으로 알려져 있다.[3,4] 이러한 기술적 발전과 함께 차량 탑재 인공지능 시스템의 실용화를 위해서는 탐지 성능뿐만 아니라 시스템 운용 환경에서의 물리적 제약을 함께 고려하는 것이 필수적이다.

그러나 차량에 탑재되는 임베디드 시스템은 서버급 연산 장치와 달리 제한된 전력 예산과 열 설계 한계를 가진다. 높은 연산 부하는 전력 소비 증가와 시스템 발열을 유발하며, 이는 연산 성능 저하 또는 시스템 안정성 저하로 이어질 수 있다.

Fig. 1은 차량 탑재 임베디드 시스템에서 고려되어야

*Corresponding author, E-mail: sckee@chungbuk.ac.kr

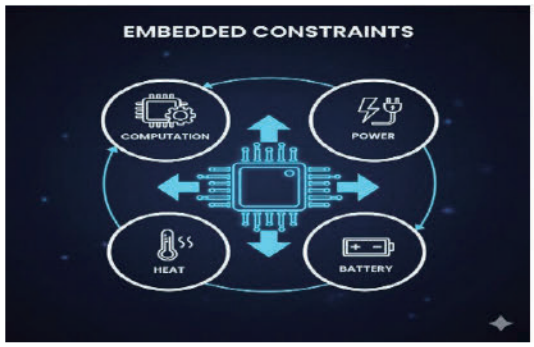

Fig. 1 Four critical constraints of embedded systems

할 네 가지 핵심 제약 요소를 개념적으로 나타낸다. 따라서 차량 탑재 객체 탐지 시스템에서는 탐지 정확도와 추론속도 뿐만 아니라 전력 소비를 포함한 에너지 효율을 종합적으로 고려한 모델 설계 및 평가가 필수적이다.[5] 그러나 실제 차량 탑재 환경에서는 전력 소비 및 에너지 효율에 대한 정량적 분석이 상대적으로 부족하며, 운용 목적에 따라 최적 모델을 선택하기 위한 체계적인 평가 프레임워크 연구 역시 제한적인 실정이다.[6-8]

특히 동일한 탐지 정확도를 유지하는 모델이라 하더라도 적용된 경량화 기법 및 하드웨어 가속 방식에 따라 전력 소비 특성은 크게 달라질 수 있다. 이러한 차이는 배터리 기반 시스템의 운용 시간, 발열로 인한 성능 저하, 시스템 안정성에 직접적인 영향을 미친다. 따라서 실제 차량 탑재 환경에서는 단일 성능 지표의 우수성보다 제한된 전력 예산과 열 제약 하에서 운용 목적에 적합한 모델을 선택하는 것이 더욱 중요하다. 이를 위해서는 탐지 정확도, 추론 속도, 전력 소비 및 에너지 효율을 통합적으로 고려하는 다중 지표 기반의 평가 체계가 요구된다.

본 논문에서는 Fig. 2에 나타낸 차량 탑재 임베디드 객체 탐지 시스템 구조를 기반으로, YOLO 기반 객체 탐지 모델의 전력 효율 관점 성능을 체계적으로 평가하고, 이를 바탕으로 운용 목적에 적합한 모델을 선택하기 위한 평가·선정 프레임워크를 제안한다. YOLOv5 기반 객체 탐지 모델을 기준으로 구조적 가지치기와 사후 학습 양자화 기법을 적용한 여러 후보 모델을 동일 조건에서 비교하고, 탐지 정확도, 추론 속도, 평균 전력 소비 및 에너지 효율 지표(mAP/W)를 통합적으로 분석한다. 특히 본 연구의 차별성은 개별 경량화 기법 자체를 새롭게 제안하는 데 있는 것이 아니라, 차량 탑재 임베디드 객체 탐지 환경을 대상으로 다중 성능 지표 정규화, 운용 시나리오 기반 가중치 설정, 제약 조건 필터링 및 Pareto 프런티어 분석을 하나의 의사결정 절차로 통합한 평가 프레임워크를 제안한다는 점에 있다. 이를 통해 단일 성능 지표 중심 평가의 한계를 극복하고, 실제 임베디드 환경의 제약

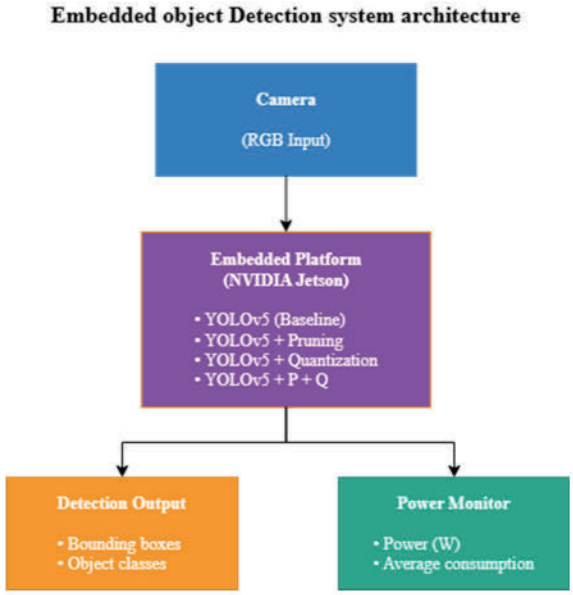

Fig. 2 Embedded object detection system architecture

조건을 반영한 목적 지향적 모델 선택을 체계적으로 지원하고자 한다.

2. 관련 연구

딥러닝 기반 객체 탐지 기술은 자율주행 차량 및 ADAS 분야에서 핵심 인식 기술로 활용되어 왔다.[1,2] 기존 객체 탐지 기법으로는 Faster R-CNN, SSD, YOLO 계열 모델 등이 제안되었으며, 이 중 YOLO 계열 모델은 단일 단계 검출 구조를 기반으로 높은 실시간성을 제공하여 임베디드 환경에 적합한 것으로 평가된다.[3,4] 최근에는 YOLOv5, YOLOv7 등 다양한 경량 객체 탐지 모델이 제안되었으며, 모델 구조 개선을 통해 연산 효율을 향상시키기 위한 연구가 지속적으로 이루어지고 있다.[9-11] 한편, 임베디드 환경에서는 제한된 연산 자원과 전력 예산으로 인해 객체 탐지 모델의 경량화가 필수적인 요소로 고려된다. 이를 위해 가지치기와 양자화와 같은 모델 압축 기법이 널리 활용되어 왔으며, 이러한 기법은 연산량 감소 및 추론 속도 향상에 효과적인 것으로 보고되었다.[6,7] 또한 최근에는 저정밀 연산, 하드웨어 가속 및 에너지 효율 중심 설계를 포함한 연구도 활발히 진행되고 있다.[8,12,13] 예를 들어, 딥러닝 연산의 에너지 소비 특성을 분석하거나, INT8 기반 저정밀 연산을 활용하여 전력 효율을 향상시키는 연구가 보고되었으며, 일부 연구에서는 하드웨어-소프트웨어 공동 최적화를 통해 에너지 효율을 개선하는 접근도 제안되고 있다.

그러나 기존 연구는 주로 개별 모델의 정확도 향상, 경량화 기법 적용, 또는 특정 하드웨어 환경에서의 성능 비

교에 초점을 맞추고 있다. 특히 차량 탑재 임베디드 환경에서 요구되는 다중 성능 지표, 즉 정확도, 실시간성, 전력 소비 및 에너지 효율을 통합적으로 고려하고, 운용 목적에 따라 최적 모델을 선택하기 위한 체계적인 프레임워크 연구는 제한적이다. 또한 일부 연구에서는 동적 모델 선택 기법이 제안되었으나, 주로 일반 CNN 추론 대상으로 수행되었으며 객체 탐지 모델에 대한 적용은 충분히 다루어지지 않았다.[14)]

따라서 본 연구에서는 이러한 한계를 보완하여, 차량 탑재 임베디드 객체 탐지 환경을 대상으로 정확도, 추론 속도, 평균 전력 소비 및 에너지 효율을 통합적으로 고려하는 모델 선택 프레임워크를 제안한다. 제안하는 프레임워크는 성능 지표 정규화, 가중 합산 기반 평가, 제약 조건 필터링 및 Pareto 프런티어 분석을 하나의 의사결정 절차로 통합함으로써, 실제 운용 목적에 적합한 객체 탐지 모델을 체계적으로 도출할 수 있도록 설계되었다.

Table 1은 기존 대표 연구와 본 연구에서 제안하는 프레임워크를 주요 평가 항목별로 비교한 것이다. 기존 연구는 주로 탐지 정확도 또는 추론 속도 중심의 단일·이중 지표 평가에 머무르고 있으며, 에너지 효율을 통합적으로 고려하거나 운용 목적에 따라 가중치를 조정할 수 있는 의사결정 구조를 제공하는 연구는 찾아보기 어렵다. 이에 반해 본 연구는 정확도, 추론 속도, 평균 전력 소비, 에너지 효율의 4가지 지표를 동시에 고려하며, 운용 시나리오 기반 가중치 설정, 제약 조건 필터링 및 Pareto 프런티어 분석을 하나의 절차로 통합함으로써 기존 연구 대비 방법론적 차별성을 제공한다.

Table 1 Comparison of prior studies and the proposed evaluation framework

Comparison item	Prior studies	Proposed framework
No. of evaluation metrics	1 ~ 2	4 (mAP, FPS, P, E)
Energy efficiency metric (mAP/W)	X	O
Multi-metric normalization	X	O
Scenario-based weight config.	X	O
Constraint-based filtering	X	O
Pareto frontier analysis	X	O
Target platform	General / Server	Vehicle-mounted
Integrated decision procedure	X	O

3. 전력 효율 기반 모델 선택 프레임워크

본 장에서는 차량 탑재 임베디드 환경에서 객체 탐지 모델을 목적 지향적으로 선택하기 위한 전력 효율 기반 모델 선택 방법을 제안한다. 기존 객체 탐지 모델 평가는 주로 탐지 정확도와 추론 속도를 중심으로 이루어져 왔으며, 모델 경량화 연구 또한 연산량 감소 및 처리 속도 향상에 초점을 맞추어 진행되어 왔다. 그러나 실제 차량 탑재 임베디드 시스템에서는 전력 소비 및 발열 제약이 시스템 안정성과 운용 시간에 직접적인 영향을 미치므로, 단일 성능 지표 기반 평가만으로는 적절한 모델 선택이 어렵다.

본 연구의 목적은 새로운 객체 탐지 모델이나 경량화 알고리즘을 제안하는 것이 아니라, 기존 모델 및 경량화 기법을 동일한 조건에서 비교하고, 다중 성능 지표를 통합하여 운용 목적에 적합한 모델을 선택할 수 있는 체계적인 의사결정 프레임워크를 제안하는 데 있다. 특히 제안 프레임워크는 성능 지표 정규화, 가중 합산 기반 평가, 제약 조건 기반 필터링 및 Pareto 프런티어 분석을 하나의 절차로 통합함으로써 기존의 단순 성능 비교를 넘어 실제 운용 환경을 반영한 모델 선택을 가능하게 한다.

아래 Fig. 3은 제안하는 전력 효율 기반 모델 선택 프

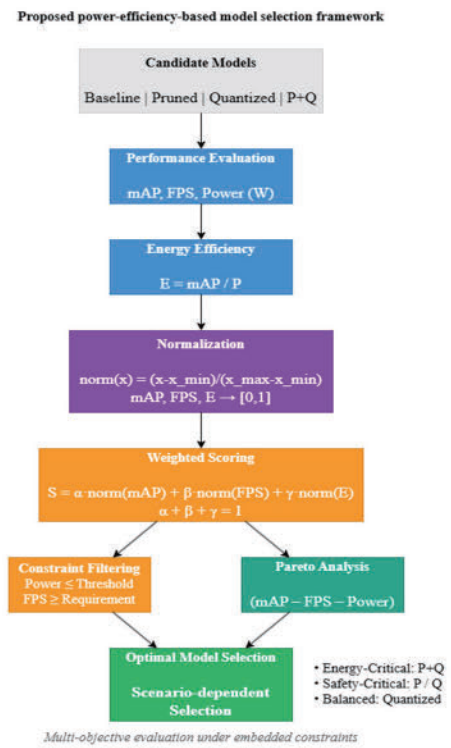

Fig. 3 Proposed power-efficiency-based model selection framework

레임워크의 전체 구조를 나타낸다. 입력 단계에서는 다양한 객체 탐지 모델 및 경량화 모델이 후보로 구성되며, 각 모델에 대해 mAP, FPS, 평균 전력 소비(W)가 측정된다. 이후 에너지 효율(mAP/W)을 산출하고, 서로 다른 단위를 갖는 성능 지표를 정규화 한다. 정규화된 지표는 가중 합산을 통해 종합 점수로 변환되며, 동시에 제약 조건 기반 필터링과 Pareto 프런티어 분석을 통해 최종 후보 모델 집합이 도출된다. 이러한 구조는 단일 성능 지표가 아닌 다중 지표 기반 의사결정을 가능하게 한다.

3.1 모델 구성 및 경량화 전략

본 연구에서는 YOLOv5 기반 객체 탐지 모델을 기준 모델(Baseline)로 설정하였다. 해당 모델은 실시간 객체 탐지에 널리 사용되는 대표적인 구조로, 다양한 임베디드 환경에서의 적용 가능성이 검증되어 있다. 다만 제안하는 프레임워크는 특정 모델 구조에 종속되지 않으며, 다양한 객체 탐지 모델 및 경량화 기법에 동일하게 적용 가능한 일반성을 가진다.

임베디드 환경에서의 연산 효율 향상을 위해 다음의 경량화 기법을 적용하였다.

1) 구조적 가지치기(Structured pruning)

중요도가 낮은 채널 또는 필터를 제거함으로써 연산량을 감소시키는 방법으로, 실제 하드웨어 연산량 감소에 직접적으로 기여한다.

2) 사후 학습 양자화(Post-training quantization)

부동소수점(FP32)연산을 정수(INT8) 연산으로 변환하여 연산 효율 및 전력 효율을 향상시킨다.[7,15]

3) 가지치기-양자화 결합 모델(P+Q)

Fig. 4는 가지치기와 양자화를 결합한 모델의 처리 절차를 나타낸다. 구조적 가지치기를 통해 모델을 축소한 후 양자화를 적용함으로써, 연산량 감소와 저정밀 연산 효과를 동시에 확보하도록 구성하였다.

3.2 성능 지표 정의

임베디드 객체 탐지 모델의 성능 평가는 단일 지표가 아닌 다중 지표 기반으로 수행된다.

1) 탐지 정확도(mAP)

객체 탐지 성능을 나타내는 대표 지표로, 무차원(Dimensionless)값으로 정의된다.

2) 추론 속도(FPS)

초당 처리 가능한 프레임 수로 정의되며, 실시간성을 평가하는 핵심 지표이다.

3) 평균 전력 소비(P)

추론 과정 동안 측정된 평균 전력 소비량(W)을 사용한다.

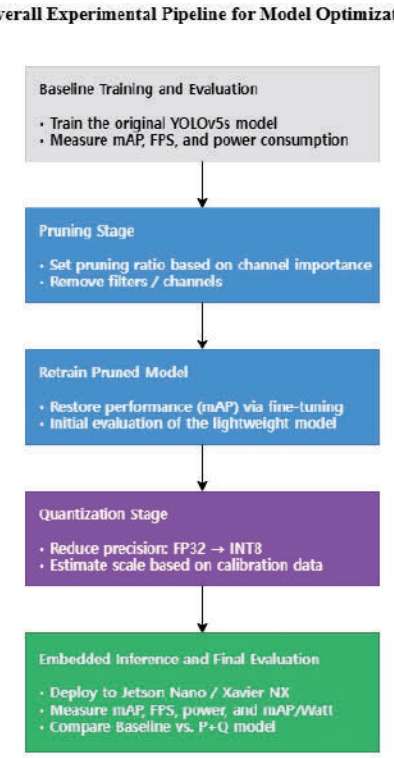

Fig. 4 Pruning + quantization procedure diagram

4) 에너지 효율 (E)

에너지 효율은 단위 전력당 탐지 성능을 나타내는 지표로 정의되며, 다음과 같이 표현된다.

$$E = mAP / P \quad (1)$$

여기서 mAP는 무차원 값으로, E는 단위 전력당 상대적 성능 지표의 의미를 가진다. 이는 일반적인 물리적 효율(%)과는 구별된다. 이러한 에너지 효율기반 평가는 최근 에너지 인지형 딥러닝 설계 및 하드웨어 가속 연구에서도 중요한 평가 기준으로 활용되고 있다.[12,13]

3.3 성능 지표 정규화

mAP, FPS, E는 서로 다른 단위와 범위를 가지므로 직접 가중 합산이 불가능하다. 이를 해결하기 위해 최소–최대 정규화(min–max normalization)를 적용하였다.

$$norm(x) = (x - x\_min) / (x\_max - x\_min) \quad (2)$$

여기서 x는 특정 모델의 성능 지표 값이며, x_min, x_max는 후보 모델 집합에서의 최소값과 최대값을 의미한다. 정규화 과정을 통해 각 지표는 0~1 범위로 변환함으로써, 다중 지표 간 공정한 비교 및 통합 평가가 가능해진다.

Table 2 Results of normalized performance metric calculations

Model	mAP	Norm (mAP)	FPS	Norm (FPS)	E	Norm (E)
Baseline	73.6	1.000	28.4	0.000	3.94	0.000
Pruned	72.9	0.300	34.7	0.348	4.79	0.280
Quantized	73.1	0.500	41.2	0.707	6.04	0.691
P+Q	72.6	0.000	46.5	1.000	6.98	1.000

Table 2는 실험에서 측정된 각 모델의 성능 지표에 대해 최소-최대 정규화를 적용한 결과를 나타낸다. mAP의 경우 모델 간 범위가 72.6~73.6으로 제한적이나, FPS 및 에너지 효율(E)은 상대적으로 넓은 범위를 가지므로 정규화를 통해 각 지표 간 단위 차이를 제거하고 공정한 가중 합산이 가능하도록 하였다. 정규화 결과, norm(FPS) 및 norm(E)에서 P+Q 모델이 1.000으로 최상위를 기록하였으며, norm(mAP)에서는 Baseline이 1.000으로 가장 높은 값을 나타냈다.

3.4 가중 합산 기반 종합 성능 점수

정규화된 성능 지표를 기반으로 종합 성능 점수 S를 다음과 같이 정의한다.

$$S = \alpha \cdot norm(mAP) + \beta \cdot norm(FPS) + \gamma \cdot norm(E) \quad (3)$$

$$\alpha + \beta + \gamma = 1 \quad (4)$$

여기서 α, β, γ는 각각 정확도, 실시간성, 에너지 효율에 대한 가중치 계수이다. 본 연구에서는 대표적인 운용 시나리오를 고려하여 가중치를 Table 3과 같이 설정하였다.

Table 3 Weight configuration for different scenarios

Scenario	α	β	γ
Accuracy-oriented	0.6	0.2	0.2
Real-time oriented	0.3	0.5	0.2
Energy-efficient	0.2	0.2	0.6

각 시나리오의 가중치는 다음과 같은 공학적 근거에 따라 설정하였다. Accuracy-oriented 시나리오에서 $\alpha = 0.6$으로 높게 설정한 것은 ISO 26262 기능 안전 기준에서 요구하는 탐지 신뢰성 최우선 원칙을 반영한 것이다. Real-time oriented 시나리오에서 $\beta = 0.5$를 적용한 것은 고속 주행 환경에서 처리 지연이 제어 응답 지연으로 직결되어 최소 30 (FPS) 이상의 실시간 처리가 필수적임을 고려한 것이다. Energy-efficient 시나리오에서 γ=0.6을 적용한 것은 배터리 기반 자율이동체 또는 장시간 운용 시스템에서 전력 예산 15 W 이하 조건과 발열 억제가 시스템 안정성에 직접적인 영향을 미침을 반영한 것이다. $\alpha + \beta + \gamma = 1$을 만족하는 범위 내에서 나머지 두 가중치는 동등한 중간 수준(0.2)으로 설정하여 단일 지표 편중을 방지하였다. 이러한 가중치 설정은 운용 목적에 따른 성능 요구사항을 반영하기 위한 것으로, 제안 프레임워크의 유연성을 제공한다.

3.5 Pareto 프런티어 기반 다목적 최적화

마지막으로 mAP–FPS–Power의 다차원 성능 공간에서 Pareto 프런티어 분석을 수행한다. Pareto 지배 관계를 기반으로 모든 성능 지표에서 동시에 열등하지 않은 후보 집합을 도출하며, 이는 단일 가중치 설정에 의존하지 않는 다목적 최적화 관점을 제공한다.[16] 또한 본 연구에서 제안한 전력 효율 기반 모델 선택 프레임워크는 성능 지표 정규화, 가중 합산 기반 평가, 제약 조건 기반 필터링, 그리고 Pareto 프런티어 분석을 하나의 통합된 의사결정 절차로 구성함으로써, 기존의 단순 성능 비교 방식과 차별화된다. 특히 각 단계는 상호 보완적으로 작용하여, 먼저 제약 조건을 통해 현실적인 후보 집합을 축소하고, 이후 가중 합산을 통해 목적 지향적인 평가를 수행하며, 마지막으로 Pareto 분석을 통해 다목적 최적 관점에서 최종 후보군을 도출하는 구조를 가진다. 이와 같은 절차를 통해 제안 방법은 차량 탑재 임베디드 환경에서 요구되는 정확도, 실시간성, 전력 소비 및 에너지 효율 간의 복합적인 관계를 체계적으로 반영할 수 있으며, 실제 운용 목적에 적합한 객체 탐지 모델을 선택하는 데 실질적인 의사결정 기준을 제공한다.

4. 실험 환경 및 결과 분석

본 장에서는 3장에서 제안한 전력 효율 기반 모델 선택 프레임워크의 타당성을 검증하기 위해 수행한 실험 환경, 조건 및 정량적 평가 결과를 제시한다. 모든 실험은 동일한 하드웨어 및 소프트웨어 환경에서 수행하여 모델 간 비교의 공정성을 확보하였다.

4.1 실험 환경

본 연구는 차량 탑재 환경을 가정한 GPU 기반 임베디드 플랫폼에서 수행되었다. 실험 플랫폼은 NVIDIA Jetson Nano를 기반으로 구성되었으며,[17,18] GPU 가속을 활용한 딥러닝 추론 환경을 구축하였다. 객체 탐지 시스템은 아래 Table 4와 같은 구조로 구성된다.

Table 4 Hardware and software environment

Category	Specification
OS	Ubuntu 20.04 LTS
Python	3.8.10
PyTorch	PyTorch 1.13.1 + CUDA 11.7
GPU	NVIDIA GeForce RTX 3090 (24 GB)
Embedded platform	Jetson Nano (4 GB RAM)
Framework	Ultralytics YOLOv5 (v7.0)

입력 영상은 실시간으로 임베디드 플랫폼에 전달되며, 각 모델은 동일한 추론 엔진 및 실행 조건에서 동작한다. 전력 소비는 플랫폼에서 제공하는 전력 모니터링 도구를 활용하여 측정하였다. 실험 중에는 초기 워밍업 구간을 제외하고 일정 시간 동안 안정 상태에서의 평균 전력 소비(W)를 사용하였다.

4.2 실험 조건

4.2.1 비교 모델 구성

본 연구에서는 Baseline, Pruned, Quantized, Pruned + Quantized(P+Q)의 네 가지 모델을 비교 대상으로 설정하였다. 모든 모델은 동일한 입력 해상도(640×640)와 동일한 데이터셋에서 평가되었다. 데이터셋 구성은 Table 5와 같다.

Table 5 Dataset composition and statistics

Category	Train	Valid	Test	Sum
Image	2,000	500	300	2,800
Object	12,430	3,015	1,798	17,243
Class	10	10	10	10

4.2.2 성능 측정 방식

성능 평가는 다음 지표를 기준으로 수행하였다.

- 탐지 정확도 : mAP@50[19)]
- 추론 속도 : FPS
- 평균 전력 소비 : (W)
- 에너지 효율 : E = mAP/P (mAP/W)

각 모델에 대해 동일한 프레임 수를 입력하여 평균 값을 산출하였다.

4.3 실험 결과

4.3.1 탐지 성능 및 추론 속도 비교

Fig. 5는 네 가지 모델의 mAP@50을 비교한 결과를 나타내며, Table 6의 수치와 동일한 경향을 시각적으로 확인할 수 있다. Baseline의 mAP@50은 73.6으로 가장 높게 나타났으나, Pruned (72.9), Quantized (73.1), P+Q (72.6) 모델 또한 1.5 % 미만의 정확도 손실을 보이며 유사한 탐지 성능을 유지하였다. 이는 구조적 가지치기 및 사후 학습 양자화 적용이 탐지 정확도를 크게 훼손하지 않으면서도 모델을 경량화 할 수 있음을 의미한다.

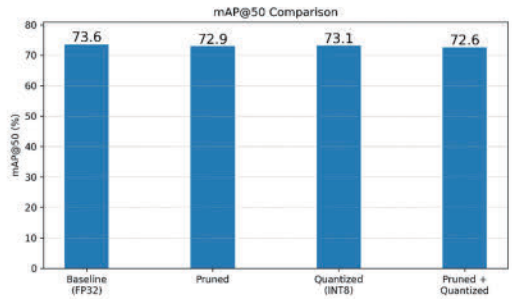

Fig. 5 mAP@50 comparison graph

Fig. 6은 모델별 FPS를 비교한 결과이며, Table 6을 통해 정량적으로 확인할 수 있다. Baseline의 FPS는 28.4로 나타난 반면, Pruned는 34.7로 약 22 % 증가하였고, Quantized는 41.21로 Baseline 대비 약 45 % 증가하였다. 특히 P+Q는 약 63 % 증가로 가장 높은 처리 속도향상을 보였으며, 이는 채널 제거를 통한 연산량 감소와 정수 기반 연산 최적화가 결합되면서 추론 처리량이 극대화된 결과로 해석된다.

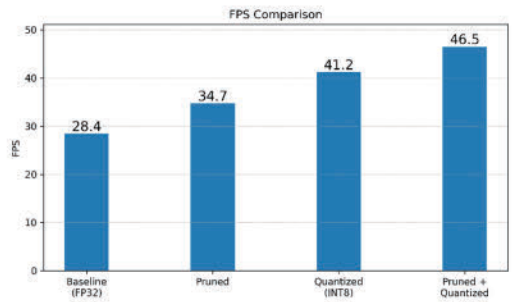

Fig. 6 FPS comparison graph

Table 6 Detection performance and inference speed comparison

Model type	Baseline	Pruned	Quantized	P+Q
mAP@50	73.6	72.9	73.1	72.6
FPS	28.4	34.7	41.2	46.5

종합하면 Figs. 5~6 및 Table 6의 결과는, 경량화 모델이 Baseline 대비 유사한 정확도를 유지하면서도 FPS를 유의미하게 향상시킬 수 있음을 보여준다. 특히 P+Q 모델은 정확도 손실이 제한적인 수준에서 가장 큰 실시간성 개선 효과를 나타냈다.

4.3.2 전력 소비 및 에너지 효율 비교

Fig. 7은 모델별 평균 전력 소비량을 비교한 결과로, Table 7의 수치와 동일한 감소 추세를 확인할 수 있다. Baseline 모델의 평균 전력 소비는 18.7 W로 가장 높았으며, Pruned는 15.2 W로 약 18 % 감소하였다. Quantized 모델은 12.1 W로 Baseline 대비 약 35 % 감소하여 전력 절감 효과가 크게 나타났고, P+Q 모델은 약 44 % 감소하며 가장 낮은 평균 전력 소비를 기록하였다. 이는 연산량 감소(Pruning)와 저정밀 정수 연산(Quantization)의 결합이 전력 효율 측면에서 상호 보완적으로 작용했음을 시사한다.

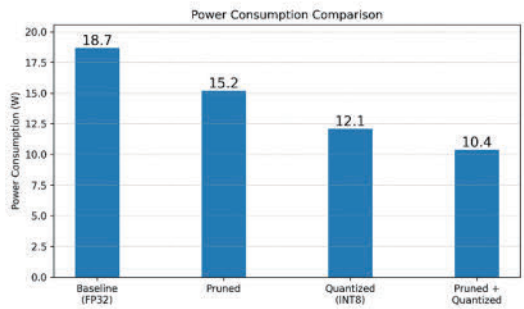

Fig. 7 Power consumption comparison graph

Fig. 8은 에너지 효율(E = mAP/W) 비교 결과이며, Table 7의 정량 결과와 일치한다. Baseline의 에너지 효율은 3.94 mAP/W로 가장 낮게 나타났고, Pruned는 4.79 mAP/W로 약 21 % 향상되었다. Quantized는 6.04 mAP/W로 약 53 % 향상되었으며, P+Q 모델은 6.98 mAP/W로 약 77 % 향상된 가장 높은 에너지 효율을 달성하였다. 이는 단위 전력당 탐지 성능 관점에서 P+Q 모델이 가장 유리함을 의미한다.

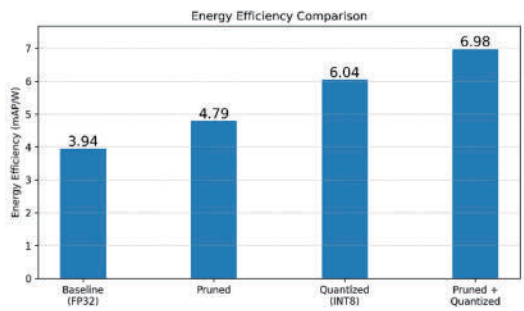

Fig. 8 Energy efficiency comparison graph

특히 Fig. 5에서 P+Q 모델의 mAP@50은 72.6으로 Baseline 대비 약 1.5 % 수준의 제한적인 감소에 그친 반면, Fig. 7에서 전력은 10.4 W로 크게 감소하여 결과적으로 Fig. 8에서 에너지 효율이 최대로 개선되었다. 따라서 Figs. 7~8 및 Table 7의 결과는 차량 탑재 임베디드 시스템에서 장시간 운용 및 발열/전력 예산 제약을 고려할 때 P+Q 모델이 가장 강력한 후보가 될 수 있음을 뒷받침한다.

Table 7 Power consumption and energy efficiency comparison

Model type	Baseline	Pruned	Quantized	P+Q
Power (W)	18.7	15.2	12.1	10.4
E (mAP/W)	3.94	4.79	6.04	6.98

이상의 Figs. 5~8 및 Tables 6~7 분석 결과를 종합하면, Pruning과 Quantization은 정확도 저하를 제한하면서 실시간성과 전력 효율을 동시에 개선할 수 있으며,[20] 특히 P+Q 모델은 FPS 및 mAP/W 측면에서 가장 우수한 균형점을 제공한다. 이는 본 연구가 제안한 전력 효율 기반 모델 선택 프레임워크의 입력 지표(mAP, FPS, E)가 실제 운용 의사결정에 유의미한 근거를 제공함을 의미한다.

4.4 프레임워크 적용 사례 분석

본 절에서는 3장에서 제안한 전력 효율 기반 모델 선택 프레임워크를 실제 실험 결과에 적용하여 시나리오별 모델 선택 결과를 분석한다. 이때 3장의 Table 2에 제시한 가중치 기준값을 사용하였다.

1) 정확도 중심 환경

탐지 누락이 치명적인 ADAS 또는 안전 중심 응용에서는 정확도 중심 가중치 $(\alpha,\ \beta,\ \gamma) = (0.6, 0.2, 0.2)$를 적용한다. 이 경우 정확도 저하가 가장 작은 모델이 상대적으로 높은 점수를 얻으며, Quantized 모델 또는 Pruned 모델이 합리적인 선택으로 도출된다.

2) 실시간성 + 효율 균형 환경

일반적인 차량 서비스 환경에서는 정확도와 에너지 효율을 유사한 수준으로 고려할 수 있으므로, 예를들어 $(\alpha,\ \beta,\ \gamma) = (0.4, 0.2, 0.4)$와 같은 균형형 가중치 설정이 가능하다. 이 경우 전력 소비와 정확도 간 균형이 우수한 Quantized 모델이 합리적인 선택으로 나타난다.

3) 에너지 효율 우선 환경

배터리 기반 드론과 같이 전력 효율이 가장 중요한 환경에서는 $\gamma > \alpha, \beta$조건에 해당하는 에너지 효율 중심 가중치 $(\alpha,\ \beta,\ \gamma) = (0.2, 0.2, 0.6)$를 적용한다. 이 경우 에너지 효율이 가장 높은 P+Q 모델이 종합 점수 S에서 최상위를 차지하면 최적 모델로 선정된다.

4) 제약 조건 기반 선택

예를 들어 평균 전력 ≤ 12 W, FPS ≥ 30과 같은 조건을 설정할 수 있다. 이 경우 해당 조건을 만족하는 모델은

Quantized와 P+Q 모델로 제한되며, 이후 종합 점수 비교를 통해 최종 모델을 선정할 수 있다.

5) Pareto 프런티어 관점

mAP–Power–FPS의 3차원 공간에서 지배 관계를 분석한 결과, P+Q 모델은 다른 모델에 의해 동시에 모든 지표에서 열등하지 않은 Pareto 우수 후보군에 포함된다. 이는 단일 지표 기반 비교를 넘어 다목적 최적화 관점에서도 경쟁력이 있음을 의미한다.

가중치 설정에 따른 프레임워크의 안정성을 검토하기 위해 Table 8과 같이 α, β, γ를 체계적으로 변화시키며 종합 점수 S의 변화를 분석하였다. 분석 결과, $\alpha \geq 0.5$인 정확도 중심 조건에서는 Baseline이, β 또는 γ가 지배적인 조건에서는 P+Q 또는 Quantized 모델이 안정적으로 최적 선택으로 도출됨을 확인하였다. 특히 $\gamma \geq 0.4$ 조건에서는 에너지 효율 지표의 상대적 우위로 인해 P+Q 모델이 일관되게 최상위를 유지하는 경향이 나타났다. 이는 제안 프레임워크가 특정 가중치 설정에 민감하게 반응하지 않으며, 운용 목적에 따라 합리적이고 일관된 모델 선택 결과를 제공함을 시사한다.

Table 8 Results of weight sensitivity analysis

α	β	γ	Baseline	Pruned	Quantized	P+Q
0.6	0.2	0.2	0.600	0.306	0.580	0.400
0.5	0.3	0.2	0.500	0.310	0.600	0.500
0.4	0.4	0.2	0.400	0.315	0.621	0.600
0.3	0.5	0.2	0.300	0.320	0.642	0.700
0.2	0.2	0.6	0.200	0.298	0.656	0.800
0.3	0.3	0.4	0.300	0.306	0.638	0.700
0.4	0.2	0.4	0.400	0.302	0.618	0.600
0.2	0.4	0.4	0.200	0.311	0.659	0.800
0.1	0.1	0.8	0.100	0.289	0.673	0.900

5. 결 론

본 논문에서는 차량 탑재 임베디드 시스템 환경에서 YOLO 기반 객체 탐지 모델의 전력 효율 관점 성능을 체계적으로 평가하였다. 기존 객체 탐지 연구가 주로 탐지 정확도와 추론 속도 향상에 초점을 맞춘 것과 달리, 본 연구는 실제 차량 적용 시 중요한 요소인 전력 소비 특성과 에너지 효율을 함께 고려하여 다양한 경량화 모델의 성능을 비교 분석하였다. YOLOv5 기반 객체 탐지 모델을 기준으로 구조적 가지치기와 사후 학습 양자화 기법을 적용한 결과, 경량화 모델은 기준 모델과 유사한 탐지 정확도를 유지하면서도 전력 소비를 유의미하게 감소시키는 것으로 나타났다. 특히 양자화 모델과 가지치기–양자화 결합 모델(P+Q)은 전력 소비 감소와 에너지 효율 향상 측면에서 우수한 성능을 보였으며, P+Q 모델은 가장 높은 에너지 효율을 달성하여 차량 탑재 임베디드 환경에서 장시간 안정적인 운용에 적합한 특성을 가지는 것으로 확인되었다. 또한 실험 결과를 통해 동일한 탐지 정확도를 갖는 모델이라 하더라도 적용된 경량화 방식에 따라 전력 소비 및 에너지 효율 특성이 크게 달라질 수 있음을 확인하였다. 이는 차량 객체 탐지 시스템 설계 시 정확도와 처리 속도뿐만 아니라 전력 효율을 핵심적인 평가 기준으로 함께 고려해야 함을 시사한다. 특히 본 연구의 의의는 개별 경량화 기법 자체를 새롭게 제안하는 데 있는 것이 아니라, 차량 탑재 임베디드 객체 탐지 환경을 대상으로 mAP, FPS, 평균 전력 소비 및 에너지 효율을 통합적으로 고려하는 평가·선정 프레임워크를 제안하였다는 점에 있다. 제안된 프레임워크는 성능 지표 정규화, 가중합산 기반 평가, 제약 조건 필터링 및 Pareto 프런티어 분석을 하나의 의사결정 절차로 통합함으로써, 단순 성능 비교를 넘어 실제 운용 목적에 적합한 모델 선택을 체계적으로 지원할 수 있음을 보여주었다. 아울러, 정확도 중심, 실시간성 중심, 에너지 효율 중심과 같은 대표적 운용 시나리오에 따라 가중치를 달리 적용할 경우 서로 다른 모델이 최적 대안으로 도출될 수 있음을 확인하였다. 이는 제안 프레임워크가 단일 기준의 우열 판단이 아니라, 운용 목적에 따라 모델 선택 기준을 유연하게 조정할 수 있는 실용적 의사결정 도구로 활용될 수 있음을 의미한다.

다만, 본 연구는 단일 임베디드 플랫폼과 제한된 데이터셋 환경에서 수행되었으며, 비교 대상 역시 YOLOv5 기반 모델과 가지치기 및 양자화 적용 모델로 제한되었다. 따라서 다양한 객체 탐지 모델 구조, 최신 경량 모델, 그리고 실제 주행 데이터 기반 환경에 대한 추가 검증이 필요하다. 향후 연구에서는 다양한 차량 운용 조건과 시스템 상태를 반영하여, 객체 탐지 모델을 실시간으로 선택하거나 전환하는 에너지 인지형 객체 탐지 시스템으로 연구를 확장할 예정이다. 이를 통해 차량 탑재 인공지능 시스템의 운용 효율과 안정성을 더욱 향상시킬 수 있을 것으로 기대된다.

후 기

본 과제(결과물)는 2025년도 교육부 및 충청북도의 재원으로 충북RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다(2025-RISE-11-003-03).

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 지역지능화혁신인재양성사업임(IITP-2026-RS-2020-II201462).

References

1) J. H. Park, S. H. Lee and K. S. Kim, "Performance Evaluation of Object Detection Algorithms for Advanced Driver Assistance Systems," Transactions of KSAE, Vol.28, No.3, pp.215-223, 2020.
2) H. Huh, "Identification of Autobody Crashworthiness for Space-Framed Vehicle Models by Finite Element Limit Analysis," Int. J. Automotive Technology, Vol.1, No.1, pp.101-112, 2000.
3) J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.779-788, 2016.
4) A. Bochkovskiy, C. Y. Wang and H. Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv Preprint arXiv:2004.10934, 2020.
5) M. J. Kim, J. W. Lee and Y. S. Cho, "Energy-Efficient Deep Learning Inference on Embedded Platforms for Automotive Applications," Transactions of KSAE, Vol.30, No.1, pp.45-53, 2022.
6) S. Han, H. Mao and W. J. Dally, "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding," International Conference on Learning Representations, 2016.
7) B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam and D. Kalenichenko, "Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.2704-2713, 2018.
8) S. Wang, Y. Li, X. Wang and Y. Qiao, "Towards Energy-Efficient Deep Neural Networks on Embedded Systems," IEEE Embedded Systems Letters, Vol.12, No.2, pp.41-44, 2020.
9) G. Jocher, A. Chaurasia and J. Qiu, YOLOv5, GitHub, https://github.com/ultralytics/yolov5, 2026.
10) A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv Preprint arXiv:1704.04861, 2017.
11) M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," Proceedings of the 36th International Conference on Machine Learning, pp.6105-6114, 2019.
12) Y. Chen, T. Krishna, J. Emer and V. Sze, "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," IEEE Journal of Solid-State Circuits, Vol.52, No.1, pp.127-138, 2017.
13) H. Zhang, Y. Chen and V. Sze, "Energy-Efficient Neural Network Design Using Precision Scaling," IEEE Transactions on Circuits and Systems for Video Technology, Vol.29, No.9, pp.2614-2628, 2019.
14) Y. Kang, J. Yoo and S. Yoo, "Dynamic Model Selection for Energy-Efficient CNN Inference on Embedded Platforms," IEEE Access, Vol.9, pp.120345-120356, 2021.
15) M. Courbariaux, Y. Bengio and J. P. David, "BinaryConnect: Training Deep Neural Networks with Binary Weights during Propagations," Advances in Neural Information Processing Systems, pp.3123-3131, 2015.
16) L. Pareto, Manuale di economia politica, Società Editrice Libraria, Milan, 1906.
17) Q. C. Cai, K. H. Lee, C. H. Lee and C. B. Hong, "Real-Time Object Detection for Autonomous Vehicles Using Embedded GPUs," KSAE Annual Conference Proceedings, pp.123-130, 2019.
18) NVIDIA Corporation, NVIDIA Jetson Platform, https://developer.nvidia.com/embedded-computing, 2025.
19) T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick and P. Dollár, "Microsoft COCO: Common Objects in Context," Proceedings of the European Conference on Computer Vision, pp.740-755, 2014.
20) V. Sze, Y. H. Chen, T. J. Yang and J. S. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," Proceedings of the IEEE, Vol.105, No.12, pp.2295-2329, 2017.