일반논문 (Regular Paper)
방송공학회논문지 제29권 제5호, 2024년 9월 (JBE Vol.29, No.5, September 2024)
https://doi.org/10.5909/JBE.2024.29.5.703
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

엣지 디바이스용 실시간 열화상 객체 검출을 위한 YOLOv5기반 경량화 방법론

윤 현 석[a)], 김 응 태[a)‡]

YOLOv5-based Lightweight Methodology for Real-time Thermal Image Object Detection for Edge Devices

Hyun-Seok Yoon[a)] and Eung-Tae Kim[a)‡]

요 약

딥러닝을 이용한 객체검출은 대부분 RGB 영상을 기반으로 이루어지고 있다. 하지만 RGB 영상은 야간이나 안개, 비, 눈 등 여러 환경적인 조건에서는 객체를 검출하는데 정확도가 떨어지는 문제점이 있다. 반면에 IR (Infra Red) 영상은 열 정보만으로 처리하므로 RGB 영상보다 해상도가 떨어지지만 환경적인 조건에 덜 민감하다. 딥러닝을 이용한 객체검출 모델은 다양하지만, 최근에는 정확도가 높고 처리 속도가 빠른 YOLO가 많이 사용되고 있다. 그러나 YOLO는 RGB 영상에 맞춰 설계되어 있어 IR 영상에서 객체검출을 하기 위해서는 IR 영상에 맞춰 설계할 필요가 있다. 또한 많은 계산량이 요구되는 딥러닝 알고리즘을 엣지 디바이스에서 사용하기 위해서는 실시간 처리를 위한 경량화가 필요하다. 본 논문은 기존 YOLOv5 모델을 IR 영상에 맞게 수정하고 경량화하기 위해 기존의 3개 레이어로 이루어진 헤드를 2개 레이어로 줄이고, 정확도 향상을 위해 CBAM (Convolution Block Attention Module)을 추가하였다. 모의실험 결과, 본 논문에서 제안한 모델의 정확도는 95.5%, mAP50은 96.4%, 파라미터는 160만 개로 기존 모델 YOLOv5s와 비교해 보면 정확도는 1.1% 떨어지지만, 파라미터는 4.375배 줄어들었다. 또한, 제안된 모델에 TensorRT 모델로 변환한 결과 mAP50은 1.6% 감소하였지만 속도 측면에서 제안된 모델보다 25fps 높은 결과를 확인했다.

Abstract

Most object detection models based on deep learning are primarily designed for RGB images. However, they suffer from poor accuracy in various environmental conditions such as nighttime, fog, rain, and snow. On the other hand, IR (Infrared) images are processed with only thermal information, so the resolution is lower than RGB images, but they are less sensitive to environmental conditions. While there are various object detection models using deep learning recently, YOLO, known for its high accuracy and fast processing speed, has been widely adopted. However, since YOLO is designed based on RGB images, it needs to be adapted for IR images to perform object detection effectively. In addition, in order to use deep learning algorithms that require a large amount of computation in edge devices, it is necessary to reduce weight for real-time processing. In this paper, in order to modify and reduce the weight of the existing YOLOv5 model to fit the IR image, the existing three-layer head was reduced to two layers, and CBAM was added to improve accuracy. In experimental results,, the precision of the model proposed in this paper was 95.5%, mAP50 was 96.4%, and the parameters were 1.6M. Compared to the existing model YOLOv5s, the precision decreased by 2.4%, but the parameters decreased by 4.375 times. In addition, as a result of converting the proposed model to the TensorRT model, mAP50 decreased by 1.6%, but in terms of speed, the result was 25fps higher than the proposed model.

Keyword : Deep Learning, Object Detection, Infrared Radiation, Attention, Edge Device, Lightweight

Ⅰ. 서 론

최근까지 딥러닝 영상처리는 초해상도(Super Resolution), 생성모델(GAN), 이미지 분할(Image Sementation), 객체 검출(Object Detection) 등 다양한 분야에서 사용되어지고 있다. 이 중 객체 검출은 자율 주행 자동차, 드론, CCTV 등에 활용되면서 엣지 디바이스에서의 실시간 객체 검출에 관한 많은 연구가 이루어지고 있다.

딥러닝을 이용한 객체 검출은 CNN(Convolutional Neural Network)[1]을 활용하여 만든 모델을 사용하여 이미지 내에서 객체가 존재할 가능성이 있는 위치를 찾는 물체 감지(Object Localization), 물체가 감지되면 해당 물체가 어떤 종류의 물체인지 분류하는 물체 분류(Object Classification)를 수행하는 작업이다. Alexnet[2]을 시작으로 꾸준히 발전해 오면서 최근에는 정확도가 높고 속도가 빠른 YOLO[3]를 사용한다.

대부분의 딥러닝 객체검출은 RGB 영상에 맞추어 설계되어 있다. 하지만 RGB 영상은 안개, 비, 눈, 빛 번짐 등 환경적인 조건에 취약하다. 이에 반해, IR 영상은 열 정보만으로 처리하므로 환경적인 조건에 유리하다. 따라서 IR 영상을 이용한 딥러닝 객체검출을 수행하려면 IR 영상에 맞게 모델을 설계해야 할 필요가 있다.

IR 영상기반 객체 검출은 꾸준히 연구가 이루어지고 있으며 공간 필터링, 주파수 필터링과 같은 방법으로 노이즈를 제거하는 방법을 사용하여 객체를 감지하는데 성능을 높이는 여러 연구가 있었다. 그러나 이 방법은 단순히 노이즈만 제거 하는 방법으로, RGB 영상 대비 해상도가 낮은 IR 영상은 객체가 더 작은 객체로 출력되기 때문에 소형 객체를 탐지하는 기법이 필요하다. 또 다른 연구로는 저해상도 영상을 고해상도로 업스케일링 해주는 초해상도 기법도 있지만 초해상도와 객체 검출을 순차적으로 진행하면 모델이 무거워 실시간으로 처리하기에는 역부족이다.

딥러닝 객체검출은 많은 연산량이 필요함에 따라서 보통 클라우드 컴퓨팅을 이용하나 최근에는 제품 내 내재된 마이크로프로세서를 이용한 엣지 디바이스에서 처리하고자 하는 필요성이 대두되고 있다. 그러므로 엣지 디바이스에서 실시간으로 처리하기 위해서는 모델의 경량화가 필요하다. 경량화 기법의 종류는 크게 가지치기, 양자화, 지식 증류, 경량 네트워크 설계가 있다. 본 논문에서는 위에 언급했듯이 IR 영상에 맞게 모델을 설계하여야 하므로 IR 영상 모델에 맞는 경량 네트워크 설계를 하고자 한다.

모델의 경량화를 하면 속도는 빨라지지만 정확도는 떨어진다. 이를 위해 객체 검출 성능을 높이는 기법인 어텐션 메카니즘인 CBAM[4]을 개선 적용한다. 본 논문은 엣지 디바이스에서 실시간 열화상 객체 검출을 하기 위해 모델을 IR 영상에 맞게 설계 및 경량화를 하고 정확도 유지를 위해 CBAM을 적용한 모델을 제안한다.

본 논문은 다음과 같이 구성된다. 본 논문의 2장에서는 관련 이론을 소개하고 3장에서는 제안된 모델의 구조를 설명한다. 4장에서는 기존의 모델과 제안된 모델을 비교하고 성능을 평가하며 5장에서 결론을 맺는다.

Ⅱ. 기존 객체 검출 관련 연구

1. 전통적인 열화상 객체 검출

전통적인 열화상 객체 검출 기법[5]으로는 히트맵, 임계값 처리, 엣지 검출 등 다양한 방법이 있다. 히트맵 기반 객체 검출을 보면, 객체는 일반적으로 열화상 영상에서 높은 온도를 가지므로, 이를 감지하기 위해 히트맵을 사용한다. 하지만 히트맵 기반의 객체 검출은 객체의 크기가 작거나 뭉게져 있을 경우 정확한 위치를 찾기가 어렵고 배경에서 발생하는 높은 온도나 노이즈에 민감하여 객체를 식별하기 어려울 수 있다. 임계값 처리 기반 객체 검출은 특정 임계값 이상의 온도를 가지는 픽셀을 객체로 간주하는 것이다. 하지만 이 방법은 환경이나 조명, 그리고 주위 온도에 민감하여 임계값을 설정하는 것이 어려울 수 있다. 마지막으로 엣

a) 한국공학대학교(구 한국산업기술대학교) 전자공학부(Department of Electronics Engineering, Tech University of KOREA)
‡ Corresponding Author : 김응태(Eung Tae Kim)
E-mail: etkim@tukorea.ac.kr
Tel: +82-31-8041-0488
ORCID: http://orcid.org/0000-0001-5984-0045
※ This work was supported by project for Industry-University-Research Institute platform cooperation R&D funded by the Ministry of SMEs and Startups in 2022.(S3312736)
· Manuscript August 7, 2024; Revised September 8, 2024; Accepted September 9, 2024.

지 검출은 객체의 경계나 모서리가 일반적으로 열화상 영상에서 변화가 큰 지점으로, 엣지 검출 알고리즘을 사용하여 객체의 특징점을 찾을 수 있다. 하지만 이 방법은 영상 내의 노이즈가 민감하여 작은 잡음이나 세세한 변화도 엣지로 감지될 수 있다. 또한, 엣지 검출은 텍스처의 세부 정보를 놓칠 수 있다. 즉, 텍스처의 작은 세부 사항은 엣지로 감지되지 않아 객체 검출을 하는데 어려울 수 있다.

2. 딥러닝 기반 객체 검출

객체 검출 분야와 관련하여 합성 곱 신경망을 사용하여 많은 알고리즘들이 제안되었다. 객체 검출 기법은 크게 2-stage detector와 1-stage detector로 나뉜다. 2-stage detector는 지역 제안(Regional Proposal)과 분류(Classification)가 순차적으로 이루어진다. 첫 번째 단계에서는 지역 제안으로 이미지 내에서 객체가 있을 것으로 예상하는 지역을 찾는 것이다. 이러한 지역을 후보 영역으로 제안하며, 주로 사용하는 방법으로는 Selective Search[6], Edge Boxes[7], 또는 Region Proposal Network[8]와 같은 기술을 사용한다. 두 번째 단계에서는 첫 번째 단계에서 생성된 후보영역에서 객체의 위치를 정확하게 파악하고 클래스를 분류한다. 이를 위해 CNN 기반의 딥러닝 모델을 사용하여 후보 영역에서의 객체의 특징을 추출하고, 객체의 위치와 클래스를 예측한다. 2-stage detector는 순차적으로 진행되기 때문에 높은 정확도를 가지지만 모델 구조가 복잡하여 설정이 더 복잡하고 처리속도가 느리다. 대표적인 네트워크로는 R-CNN[9], Fast R-CNN[10], Faster R-CNN 등이 있다.

1-stage detector는 2-stage detector의 느린 처리 속도의 단점을 개선하기 위해 등장한 기법이다. 1-stage detector는 지역 제안과 분류를 순차적으로 진행하는 2-stage detector와 다르게 동시에 진행한다. 1-stage detector의 대표적인 네트워크는 SSD[11], YOLO로 이전에는 정확도가 Faster R-CNN보다 낮았지만 YOLO모델의 새로운 버전이 계속 나오면서 빠른 속도로 처리할 뿐 아니라 정확도도 더 높아졌다. 그 중 YOLOv5[12]는 모델의 크기에 따라 n, s, m, l, x로 나누어진다. YOLOv5n은 모델의 크기가 작아 속도가 빠르지만 정확도가 다소 떨어지며 YOLOv5x은 모델의 크기가 제일 크고 속도가 느리지만 높은 정확도를 갖는다.

YOLOv5는 백본(Backbone), 넥(Neck), 헤드(Head)로 구성되어 있으며, 백본에서는 입력 이미지의 특징들을 추출

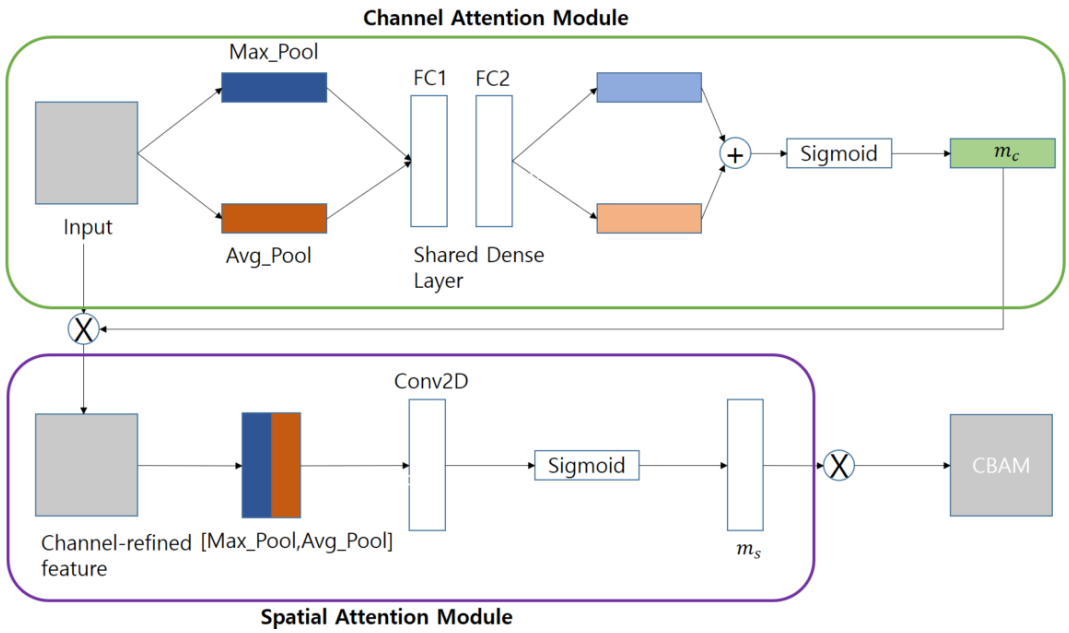


그림 1. CBAM의 구조
Fig. 1. The architecture of CBAM

하고, 넥에서는 추출된 특징을 조합하며, 헤드에서 추출한 특징을 이용하여 객체를 탐지한다.

3. 어텐션 메카니즘(Attention Mechanism)

전통적인 소형 객체 검출 방법[13]에는 영상에서 명암의 차이를 이용하여 픽셀 값을 이진화 하는 방법 및 모폴러지 연산, SIFT, SURF 등 다양한 방법이 있다. 하지만 이 방법들은 이미지의 복잡성, 다양성, 크기 변화 등에 취약하다. 이것을 보완하기 위한 방법으로 어텐션 메카니즘이 도입되었다. 어텐션 메카니즘은 검출해 내고자 하는 객체의 특징을 집중하여 원하는 정보들을 효율적으로 추출하여 객체 검출의 정확도 향상에 도움을 주는 기법이다. 어텐션 메카니즘은 채널 간의 상호 작용을 강화하고, 특정 공간적인 위치에 따라 가중치를 부여하여 작은 객체에 대한 감지 능력을 향상시킨다. 어텐션 메카니즘의 종류는 Soft Attention[14], Hard Attention[15], SE Module[16], CBAM 등 다양하다. 그 중 제일 성능이 뛰어나며 컨볼루션 레이어에 쉽게 부착할 수 있는 기법이 CBAM이다.

CBAM은 채널 어텐션(Channel Attention)과 공간 어텐션(Spatial Attention)으로 이루어져 있다. 채널 어텐션은 특징 맵의 각 채널이 얼마나 중요한지를 조절한다. 예를 들어, 특정 채널이 특정 패턴을 감지하는 데 높은 중요성을 가질 수 있으며, 이러한 중요성을 조절하여 네트워크가 중요한 특성을 강조하도록 한다. 공간 어텐션은 특징 맵 내에서 어떤 공간 영역이 중요한지를 결정한다. 즉, 네트워크가 이미지의 어느 부분에 초점을 맞춰야 하는지를 조절한다. CBAM의 구조는 그림 1과 같다.

4. 경량화 기법

딥러닝 모델은 영상분석 분야에서 좋은 성능을 보여주고 있지만 많은 메모리 공간과 연산량이 필요하여 효율이 떨어지는 문제점을 가지고 있다. 특히 모바일 기기에서 딥러닝 모델을 불러오기 위해서는 메모리에 파라미터의 값을 올려야 하는데, 메모리가 부족하여 어려운 경우가 있다. 또한, 딥러닝은 저장된 가중치를 활용해 많은 연산을 해야 하는데, 프로세서의 성능이 낮은 경우 이미지 처리에 오랜 시간이 걸릴 수 있다. 따라서 파라미터 수, 연산량, 처리 시간을 줄여주는 경량화 기술이 필요하다. 모델 경량화 기술은 딥러닝 모델을 더 작고 가벼운 형태로 만들어 성능을 유지하면서도 모델을 배포하고 실행하는 데 필요한 계산 및 메모리 자원을 최적화하는 프로세스를 의미한다. 모델 경량화 기술은 크게 가지치기(Pruning), 양자화(Quantization), 지식 증류(Knowledge Distillation), 경량 네트워크 설계(Compact Network Desigon)가 있다.

Ⅲ. 제안된 시스템

1. 경량화된 YOLOv5 모델

본 논문에서는 실내 환경에서 수집된 IR 영상 데이터셋을 사용하였고 물체의 비율을 분석하였다. 각 바운딩 박스의 면적을 구하고, 이 면적을 전체 이미지 면적으로 나눈 값을 구해 평균을 구해보니 0.0566의 비율이 나타났다. 0.0556은 바운딩 박스가 전체 이미지에서 차지하는 평균적

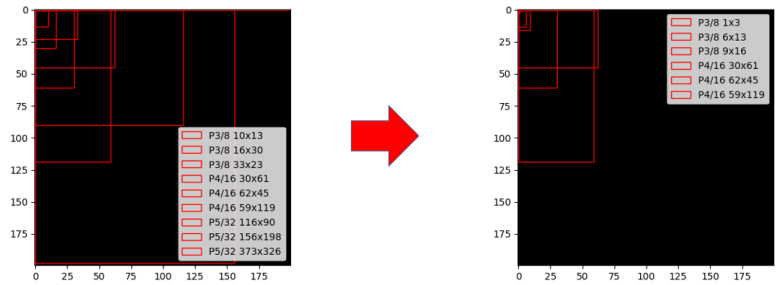


그림 2. 개선된 Anchor Box
Fig. 2. Improved Anchor Box

인 면적 비율로 이미지에서 한 물체의 바운딩 박스가 평균적으로 이미지의 약 5.66%를 차지하고 있다는 뜻이다. 이 값이 작다는 것은 데이터셋 내의 물체가 전체 이미지에 비해 상대적으로 작은 크기를 가진다는 것을 의미한다. IR 영상은 열 정보를 기반으로 처리되기 때문에, RGB 영상에 비해 멀리 있는 물체를 인식하는 데 불리한 특성을 갖고 있다. 특히, 실내 환경에서는 물체의 크기가 다양하지 않고 제한된 크기의 물체가 존재하며 물체의 크기가 대부분 중간 크기 이하이기 때문에, 큰 물체를 검출하는 것보다 더 작은 물체를 검출하는 데 집중하여야 한다. 따라서, YOLOv5 모델에서 큰 물체를 검출하는 레이어를 제거함으로써, 연산 효율성을 높이고 속도를 증가시킬 수 있고 이런 최적화는 실내 환경에서의 IR 영상의 특성과 데이터셋 분석 결과를 반영한 것이다. 이에 따라 앵커 박스 크기도 그림 2와 같이 수정하였다.

YOLOv5에서 큰 물체를 탐지하는 레이어를 삭제하면서 해당 레이어와 연결된 업샘플링 및 C3 모듈이 제거되었다. 백본에서는 SPPF 모듈의 채널 수가 1024에서 512로 줄어들었고, 최종 출력 레이어의 채널 수도 1024에서 256으로 감소했다. 이로 인해 전체 파라미터 수는 7M에서 1.6M으로 크게 줄어들었다.

2. 개선된 CBAM (Improved-CBAM)

기존의 CBAM은 채널 어텐션과 공간 어텐션으로 구성되어 있으며 순차적으로 이루어진다. 채널 어텐션은 평균 풀링(average pooling)과 맥스 풀링(max pooling)을 동시에 진행한다. 열화상 객체 검출에서 CBAM을 적용할 때 주의할 사항이 있다. 평균 풀링은 특징 맵의 평균 값을 취하므로, 이미지 전체에서 고르게 정보를 반영하여 노이즈에 덜 민감하다. 또한, 맵의 모든 값을 고려하여 평균을 내므로, 중요한 특징이 손실될 가능성이 줄어든다. 하지만 맥스 풀링은 특정 위치에서 최대 온도 값만을 반영한다. 이는 높은 온도 값이 노이즈일 때 잘못된 감지를 초래할 수 있다. 또한, 특정 위치의 정보만 반영하므로, 작은 영역의 높은 온도 값을 지나치게 강조하게 되어, 불필요한 객체를 사람으로 오인할 가능성이 높다. 따라서 본 논문에서는 채널 어텐션을 사용할 때 노이즈에 덜 민감하고 다양한 조건에서도 일관된 특징을 제공하는 평균 풀링만을 사용한다. 다만 공간 어텐션에서는 특정 위치의 최대 활성화를 포착하여 공간적인 중요한 정보를 반영하기 위해 맥스 풀링 그대로 적용한다. 개선된 채널 어텐션의 구조는 그림 3과 같다.

$$M_c(F) = \sigma(MLP(AvgPool(F))) = \sigma(W_1(W_0(F_{avg}^c))) \quad (1)$$

여기서 F는 입력 특징 맵(feature map)을 뜻하며 C개의 채널과 $H \times W$의 공간 차원을 가진다. F_{avg}^c은 입력된 특징 맵의 평균 풀링을 통해 출력된 값을 의미하며 W_0와 W_1은 채널 중요도 맵인 M_c를 계산할 때 사용하는 가중치 매트릭스이다. W_0는 첫 번째 완전 연결층의 가중치이며 $W_0 \in R^{\frac{C}{r} \times C}$로 표현할 수 있다. W_1는 두 번째 완전 연결층의 가중치이며 $W_1 \in R^{C \times \frac{C}{r}}$로 표현할 수 있다. σ는 시그모이드 함수를 뜻하며 MLP를 통해 W_0와 W_1를 공유한다.

본 논문에서는 평균 풀링만을 사용하는 I-CBAM을 헤드 마지막 출력 부분에 추가하였다. 헤드는 최종적으로 물체를 검출하는 부분으로 I-CBAM을 추가해 상대적으로 작은 물체에 집중하여 객체의 성능을 높이는데 기여하였다.

3. TensorRT를 사용한 QAT (Quantization Aware Training)

양자화는 학습이 완료된 모델의 가중치와 활성화를 32비

그림 3. 개선된 Channel Attention
Fig. 3. The architecture of Improved Channel Attention

그림 4. TensoRT 모델 변환 과정
Fig. 4. The conversion Process of TensorRT Model

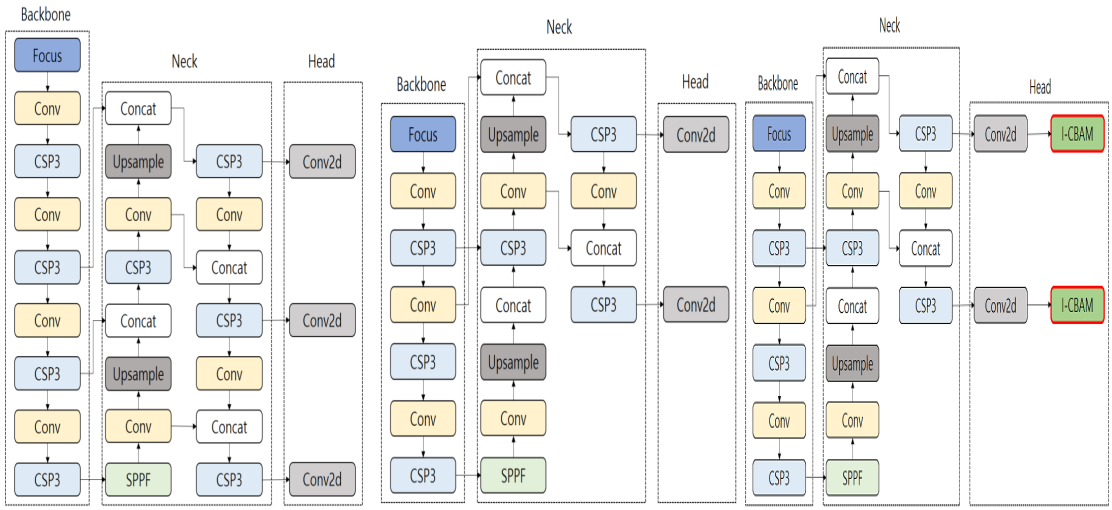


그림 5. 모델 구조 비교
Fig. 5. Model architecture comparsion

트 부동 소수점에서 8비트 정수 등으로 변환하는 과정이다. 이 과정을 통해 모델의 크기를 줄이고, 메모리 사용을 감소시키며, 계산 속도를 높일 수 있다. 양자화의 한 기법인 QAT[17]는 모델을 학습하는 동안 양자화의 효과를 시뮬레이션하는 방법이다. 즉, 모델 학습 과정에서 모든 가중치와 활성 함수에 가짜 양자화를 삽입한다. 이는 학습 후 양자화하는 방법보다 높은 정확도를 제공한다. QAT의 과정은 다음과 같다. 먼저 모델을 일반적인 방법으로 학습시켜 초기 가중치를 얻는다. 다음으로 QAT를 통한 파인 튜닝(Fine-tuning)으로 초기 학습이 완료된 모델의 가중치를 사용하여 QAT를 수행한다. 이 단계에서는 가중치와 활성화 함수에 가짜 양자화를 삽입하여 모델을 다시 학습시킨다. QAT가 완료된 모델을 ONNX (Open Neural Network Exchange) 형식으로 변환하여 .onnx 파일을 생성한다.

TensorRT란 학습된 딥러닝 모델을 최적화하여 NVIDIA GPU 상에서의 추론 속도를 수배~수십 배까지 향상시켜주는 모델 최적화 엔진이다. TensorRT는 모델 최적화, 추론 속도 향상, 메모리 효율성, 양자화 등 다양한 기능을 제공한다. 본 논문에서는 추론 속도를 향상시키기 위해 QAT가 완료된 .onnx파일을 TensorRT 모델로 변환한다.

그림 5는 각 모델 구조를 비교한 그림이고 (a)는 YOLOv5, (b)는 2scale로 경량화한 YOLOv5, (c)는 본 논문에서 제안한 구조이다.

Ⅳ. 모의실험 결과

제안된 네트워크의 성능을 확인하기 위해 데이터 셋을 직접 구성하였다. 실내 환경에서 촬영하였고 640 x 480의 크기를 가지는 IR 영상으로 이루어졌다. 클래스는 1 클래스로 “Person”으로 이루어졌다. 데이터 셋은 최소 1m부터 최대 30m까지 촬영하였으며 Train 590장, Validation 118장 Test 60장으로 총 768장으로 구성되었다. 그림 6은 건물내부의 복도, 사무실 등을 촬영한 실내 환경 예시이며 그림 7

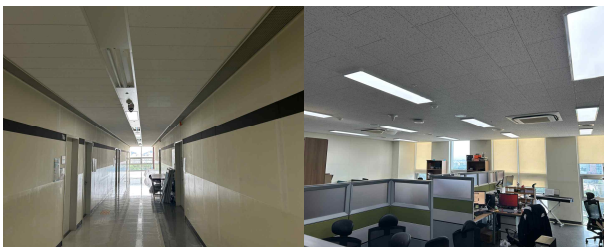


그림 6. 데이터셋 수집 환경
Fig. 6. Dataset collection environment

그림 7. 데이터셋 예시
Fig. 7. Dataset example

그림 8. QuntumRed VR
Fig. 8. QuntumRed VR

은 그림 8의 IR카메라로 찍은 데이터 셋 예시이다.

배치 값은 학습 중 입력을 동시에 몇 장을 처리할지 정하는 것으로 값이 높을수록 빠른 속도를 낼 수 있는 반면 실험 환경에 비해 너무 높은 값일 경우 메모리 부족으로 프로그램이 구동되지 않을 수 있다. 본 실험에서는 배치 값은 4로 설정하였다. 또한, 기존 앵커 박스인 [10,13, 16,30, 33,23], [30,61, 62,45, 59,119], [116,90, 156,198, 372,326]에서 큰 물체를 찾는 부분을 삭제하고 작은 물체를 찾는 부분을 수정하여 [1,3, 6,13, 9,16], [30,61, 62,45, 59,119]로 설정하였다.

표 1. QuntumRed VR 사양
Table 1. QunTumRed VR Specifications

Specification	Detail
Array Format	275 TOPS
Resolution	640 x 480
Temperature Range	~+500 $^{\circ}C$
Usage	Dataset & Test Camera

본 논문에서 제안하는 시스템은 리눅스 기반의 Ubuntu 18.04 LTS Server/Intel Xeon Gold 640R CPU @ 2.40HBz/GeForce RTX 3090 24GB 환경에서 학습을 시킨 후 열화상 카메라인 Quantemred VR을 엣지 디바이스인 Nvidia Jetson AGX Orin 개발 키트에 포팅하여 실험을 진행하였다. Quantemred VR은 그림 8과 표 1과 같고 Nvidia Jetson Orin AGX 개발 키트는 그림 9와 표 2와 같다.

그림 9. NVIDIA Jetson Orin AGX 개발 키트
Fig. 9. NVIDIA Jetson Orin AGX Developer Kit

표 2. NVIDIA Jetson Orin AGX Developer Kit 사양
Table 2. NVIDIA Jetson Orin AGX Developer Kit Specifications

Module	Detail
AI Performance	275 TOPS
GPU	NVIDIA Ampere architecture with 2048 NVIDIA CUDA cores and 64 Tensors cores
CPU	12-core Arm Cortex-A78AE v8.2 64-bit CPU 3MB L2 + 6MB L3
Memory	32GB 256-bit LPDDR5 204.8 GB/s
Usage	Edge device

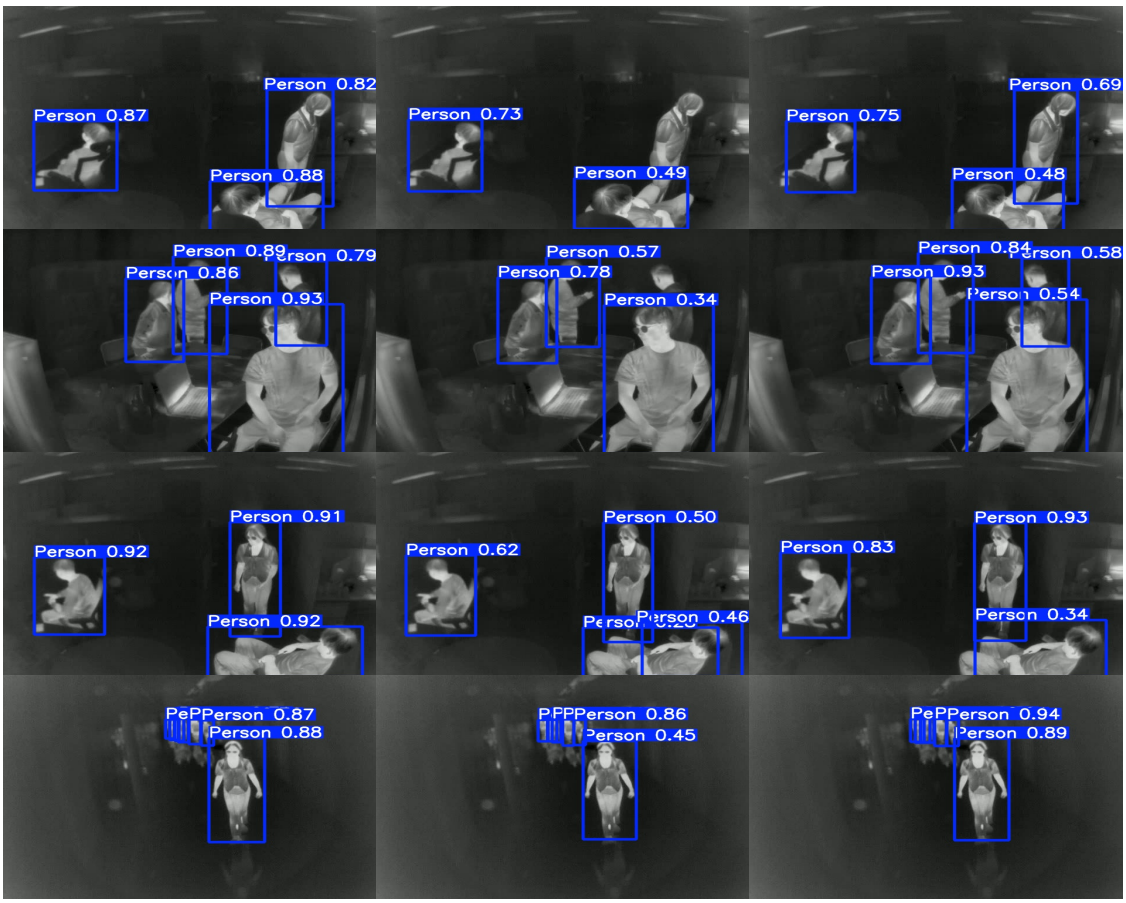
본 논문에서 사용한 모델은 모델의 크기가 작으면서 충분한 성능을 가져오는 YOLOv5s 모델을 선택했다.

본 논문에서의 첫 번째 실험에서는 기존 모델인 YOLOv5s, 2scale로 경량화한 YOLOv5s, 그리고 제안된 모델을 비교하였으며 비교 항목은 정확도, mAP50, 파라미

터 수, 전처리 시간, FLOPs로 구성되었다. 표 3에 따르면 기존 YOLOv5s와 2scale로 경량화한 YOLOv5s를 비교했을 때, 정확도는 기존 모델보다 2.4% 낮아졌지만 파라미터 수는 7M에서 1.5M으로 크게 감소하였다. 파라미터가 크게 줄어든 것에 비해 정확도가 2.4%밖에 떨어지지 않았다는 점은 큰 레이어를 삭제한 결정이 타당하다는 근거가 된다. 또한, 2scale로 경량화한 YOLOv5s에 I-CBAM을 추가한 모델은 가까운 거리에 있는 물체나 겹쳐있는 물체를 더 효과적으로 검출하였으며, 경량화한 모델에 비해 파라미터는 0.1M 증가하였지만 정확도는 1.3% 증가하였다. 늘어난 파라미터 수에 비해 정확도가 1.3% 증가한 것은 I-CBAM이 좋은 성능을 낸 것으로 볼 수 있다. 최종적으로 제안된 모델과 기존 YOLOv5s를 비교했을 때 정확도는 1.1% 낮아졌지만 파라미터는 4.375배 감소하였고 전처리 시간은 8.091ms

표 3. 기존모델과 제안된 모델의 성능 비교
Table 3. Performance comparison of existing model and proposed model

	Precision	mAP50	Parameter	Pre-processing	FLOPs
YOLOv5s	96.6	98.6	7M	32.572 m/s	16.5
2scale YOLOv5s	94.2	95.5	1.5M	22.325 m/s	12.5
Proposed Model	95.5	96.4	1.6M	24.481 m/s	12.6



(a) YOLOv5s (b) 2scale YOLOv5s (c) Proposed Model

그림 10. 모델 비교 결과
Fig. 10. Model comparison result

줄어들었다.

본 논문에서의 두 번째 실험 비교는 제안된 모델과 QAT를 적용한 후 TensorRT 모델로 변환한 모델과 비교하였다. 표 4에서 보는 바와 같이 속도 측면에서는 제안 모델은 33fps, TensorRT 모델로 변환한 모델은 58fps로 약 1.76배 증가하였다. 하지만 mAP 측면에서는 mAP50이 제안 모델은 96.4%, TensorRT 모델로 변환한 모델은 94.8%로 1.6% 감소하였다.

표 4. 제안된 모델과 TensorRT 모델의 성능 비교
Table 4. Performance comparison of proposed model and TensorRT model

Model	FPS	mAP50
Proposed Model	33fps	96.4
Proposed Model + TensorRT	58fps	94.8

그림 10은 기존 모델 (a), 2scale로 경량화한 모델 (b), 제안된 모델 (c)의 비교 결과 사진이다. 세 모델 모두 먼 거리에 있는 물체의 객체 검출 성능은 좋으나 (b)의 경우 큰 물체 및 겹쳐 있는 물체의 검출 성능이 낮은 걸 볼 수 있다. 2scale 경량화한 모델에 I-CBAM을 적용한 (c)의 결과를 보면 큰 물체 및 겹쳐있는 물체를 잘 검출하는 것을 볼 수 있다. 최종적으로 기존 YOLOv5s 모델과 비교해보면 성능에 큰 차이가 없는 것으로 보여진다.

V. 결 론

딥러닝을 이용한 객체 검출은 꾸준히 연구되면서 스마트 시티, 자율 주행 자동차, CCTV 등 많은 실생활 속에서 접목되고 있다. 컴퓨터 비전에는 많은 분야가 있는데 그 중 객체 검출은 대부분 RGB 영상에 맞춰있다. 하지만 RGB 영상은 야간 등 환경적 조건에 민감하여 IR 영상으로 대체할 필요가 있다. 또한 딥러닝 기반 객체 검출은 많은 연산량이 필요함에 따라서 보통 클라우드 컴퓨팅을 이용하나, 최근에는 제품 내 내재된 마이크로프로세서를 이용한 엣지 디바이스에서 처리하고자하는 필요성이 대두되고 있다. 그러므로 엣지 디바이스에서 실시간으로 처리하기 위해서는 모델의 경량화가 필요하다.

본 논문에서는 정확도가 높고 속도가 빠른 YOLOv5 모델을 IR 영상에 맞게 설계 및 경량화를 하였다. 기존 YOLOv5 모델에서 3스케일의 헤드를 2스케일로 수정하여 IR 영상에 맞게 설계 및 경량화를 하였고 성능 향상을 위한 개선된 CBAM을 추가하였다. 그 결과 정확도는 95.5%, mAP50은 96.4%, 파라미터는 1.6M의 결과를 얻었다. 기존 모델과 비교하면 정확도는 1.1% 감소했지만 파라미터는 4.375배 감소하였고 전처리 시간은 8.091m/s 감소하였다.

제안된 모델에 QAT 적용 후 TensorRT 모델로 변환했을 때 mAP50은 1.6% 감소했지만 fps는 제안된 모델보다 25fps 더 높은 결과를 확인하였다. 제안된 방식은 엣지 디바이스에서 열화상 영상기반 객체 검출에 빠른 프로세싱 시간과 적은 파라미터들의 사용으로 효율적임을 확인할 수 있었다.

참 고 문 헌 (References)

[1] O'shea, Keiron, and Ryan Nash. “An introduction to convolutional neural networks.” arXiv preprint arXiv:1511.08458, 2015. doi: https://doi.org/10.48550/arXiv.1511.08458
[2] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. “Imagenet classification with deep convolutional neural networks.” Advances in neural information processing systems25, 2012.
[3] Redmon, Joseph, et al. “You only look once: Unified, real-time object detection.” Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788, 2016. doi:https://doi.org/10.48550/arXiv.1506.02640
[4] Woo, Sanghyun, et al. “Cbam: Convolutional block attention module.” Proceedings of the European conference on computer vision (ECCV), pp.3-19, 2018. doi: https://doi.org/10.48550/arXiv.1807.06521
[5] Kaiwen Duan and et al, “Centernet: Keypoint tripletsfor object detection”, In P roceedings of theIEEE/CVF international conference on computervision, pp. 6569-6578, 2019. doi: https://doi.org/10.48550/arXiv.1904.08189
[6] Uijlings, Jasper RR, et al. “Selective search for object recognition.” International journal of computer vision 104, pp.154-171, 2013. doi: https://doi.org/10.1007/s11263-013-0620-5
[7] Zitnick, C. Lawrence, and Piotr Dollár. “Edge boxes: Locating object proposals from edges.” Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014. doi: https://doi.org/10.1007/978-3-319-10602-1_26
[8] Ren, Shaoqing, et al. “Faster r-cnn: Towards real-time object detection with region proposal networks.” Advances in neural information

processing systems28, 2015.
doi: https://doi.org/10.48550/arXiv.1506.01497

[9] Girshick, Ross, et al. “Rich feature hierarchies for accurate object detection and semantic segmentation.” Proceedings of the IEEE conference on computer vision and pattern recognition, pp.580-587, 2014.
doi: https://doi.org/10.48550/arXiv.1311.2524

[10] Girshick, Ross. “Fast r-cnn.” Proceedings of the IEEE international conference on computer vision, pp.1440-1448, 2015.
doi: https://doi.org/10.48550/arXiv.1504.08083

[11] Liu, Wei, et al. “Ssd: Single shot multibox detector.” Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, pp.21-37, 2016.
doi: https://doi.org/10.1007/978-3-319-46448-0_2

[12] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012,Ayush Chaurasia, TaoXie, Liu Changyu, Abhiram V, Laughing, tkianai, yxNONG, Adam Hogan, lorenzomammana,AlexWang1900, Jan Hajek, Laurentiu Diaconu, Marc,Yonghye Kwon, oleg, wanghaoyang 0106, Yann Defretin,Aditya Lohia, ml5ah, Ben Milanko, Benjamin Fineran,Daniel Khromov, Ding Yiwei, Doug, Durgesh, and FranciscoIngham. ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations, Apr. 2021. (https://github.com/ultralytics/yolov5)

[13] Cheng, Gong, et al. “Towards large-scale small object detection: Survey and benchmarks.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.
doi: https://doi.org/10.1109/TPAMI.2023.3290594

[14] Datta, Soumyya Kanti, et al. “Soft attention improves skin cancer classification performance.” Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data: 4th International Workshop, iMIMIC 2021, and 1st International Workshop, TDA4MedicalData 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 4. Springer International Publishing, pp.13-23, 2021.
doi: https://doi.org/10.48550/arXiv.2105.03358

[15] Papadopoulos, Athanasios, Pawel Korus, and Nasir Memon. “Hard-attention for scalable image classification.” Advances in Neural Information Processing Systems34, pp.14694-147-7, 2021.
doi: https://doi.org/10.48550/arXiv.2102.10212

[16] Hu, Jie, Li Shen, and Gang Sun. “Squeeze-and-excitation networks.” Proceedings of the IEEE conference on computer vision and pattern recognition, pp.7132-7131, 2018.
doi: https://doi.org/10.48550/arXiv.1709.01507

[17] Jacob, Benoit, et al. “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.
doi: https://doi.org/10.48550/arXiv.1712.05877

저 자 소 개

윤 현 석

- 2022년 : 한국공학대학교 전자공학부 공학사
- 2024년 : 한국공학대학교 전자공학구 공학석사
- ORCID : https://ordic.org/0009-0004-8568-4700
- 주관심분야 : 딥러닝기반 영상처리, 경량화, 컴퓨터 비전, 임베디드 시스템

김 응 태

- 1991년 : 인하대학교 전자공학과 공학사
- 1993년 : KAIST 전기및전자공학과 공학석사
- 1999년 : KAIST 전기및전자공학과 공학박사
- 1998년 3월 ~ 2004년 2월 : ㈜LG전자 DTV연구소 책임연구원
- 2004년 3월 ~ 현재 : 한국공학대학교(구: 한국산업기술대학교) 전자공학부 교수
- ORCID : https://orcid.org/0000-0001-5984-0045
- 주관심분야 : 멀티미디어 신호처리, DTV SOC, 지능형 영상감시 시스템, 딥러닝기반 영상처리