일반논문 (Regular Paper)
방송공학회논문지 제25권 제6호, 2020년 11월 (JBE Vol. 25, No. 6, November 2020)
https://doi.org/10.5909/JBE.2020.25.6.944
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

주의 모듈 기반 Mask R-CNN 경량화 모델을 이용한 도로 환경 내 객체 검출 방법

송 민 수[a)], 김 원 준[a)‡], 장 래 영[b)], 이 용[b)], 박 민 우[b)], 이 상 환[b)], 최 명 석[b)]

Object Detection on the Road Environment Using Attention Module-based Lightweight Mask R-CNN

Minsoo Song[a)], Wonjun Kim[a)‡], Rae-Young Jang[b)], Ryong Lee[b)], Min-Woo Park[b)], Sang-Hwan Lee[b)], and Myung-seok Choi[b)]

요 약

객체 검출 알고리즘은 자율주행 시스템 구현을 위한 핵심 요소이다. 최근 심층 합성곱 신경망 (Deep Convolutional Neural Network) 기반의 영상 인식 기술이 발전함에 따라 심층 학습을 이용한 객체 검출 관련 연구들이 활발히 진행되고 있다. 본 논문에서는 객체 검출에 가장 널리 사용되고 있는 Mask R-CNN의 경량화 모델을 제안하여 도로 내 다양한 객체들의 위치와 형태를 효율적으로 예측하는 방법을 제안한다. 또한, 주의 모듈(Attention Module)을 Mask R-CNN 내 각각 다른 역할을 수행하는 신경망 계층에 적용함으로써 특징 지도를 적응적으로 재교정(Re-calibration)하여 검출 성능을 향상시킨다. 실제 주행 영상에 대한 다양한 실험 결과를 통해 제안하는 방법이 기존 방법 대비 크게 감소된 신경망 매개변수만을 이용하여 고성능 검출 성능을 유지함을 보인다.

Abstract

Object detection plays a crucial role in a self-driving system. With the advances of image recognition based on deep convolutional neural networks, researches on object detection have been actively explored. In this paper, we proposed a lightweight model of the mask R-CNN, which has been most widely used for object detection, to efficiently predict location and shape of various objects on the road environment. Furthermore, feature maps are adaptively re-calibrated to improve the detection performance by applying an attention module to the neural network layer that plays different roles within the mask R-CNN. Various experimental results for real driving scenes demonstrate that the proposed method is able to maintain the high detection performance with significantly reduced network parameters.

Keyword : object detection, deep convolutional neural networks, lightweight model, attention module, road environment

Ⅰ. 서 론

최근 인공지능과 5G 통신기술이 빠르게 발전함에 따라 자율주행 관련 기술 또한 함께 성장하고 있으며, 안정적으로 도로 상황을 파악하는 기술 개발이 활발히 진행되고 있다. 도로 내 객체 검출기술은 자율주행 시 빠르게 변화하는 도로 환경을 이해하고 교통 흐름을 신속하게 파악하여 처리하기 위한 중요한 요소 중 하나이다. 고성능 도로 객체 검출기는 자동차 간 거리 유지 시스템, 보행자 출현 시 자동 긴급제동 시스템, 차선 이탈 경보시스템 등 지능형 자율주행 시스템에 효과적으로 적용될 수 있다. 최근 LiDAR 센서를 기반으로 한 도로 객체 검출 연구는 정확한 3차원 스캔 정보를 바탕으로 높은 검출 정확도를 도출할 수 있으나, LiDAR 센서의 비싼 가격으로 인해 상용화에 많은 어려움이 있다. 또한, LiDAR 센서는 날씨(예를 들어, 눈이나 비가 오는 날씨)와 같은 환경 변수에 민감한 단점이 있다. 따라서 최근에는 심층학습 기술의 발전에 힘입어 RGB 카메라를 기반으로 획득한 칼라 영상만을 이용하여 도로 객체를 검출하는 연구가 활발히 진행되고 있다.

합성곱 신경망(Convolution Neural Network, CNN)을 이용한 다양한 객체 인식 및 검출 방법들 중에서 특히 Mask R-CNN[1]은 객체 별 분할(Instance Segmentation) 결과까지 제공함으로써 자율주행을 위한 도로 환경 객체 검출에 널리 사용되고 있다. Mask R-CNN은 먼저 영역 제안 신경망(Region Proposal Network)을 통해 영상 내 객체가 존재할 가능성이 있는 영역 후보를 생성하고 이에 대하여 잠재 특징(Latent Feature)을 추출한다. 추출된 잠재 특징은 세 가지 작업, 즉, 객체 인식(Classification), 객체 위치 예측(Bounding Box Regression) 및 영역 분할(Segmentation)에 공통으로 사용된다. 이러한 2단계(Two-stage) 신경망 구조는 복잡한 영상에서도 목표 객체의 경계를 정확하게 검출할 수 있으며 객체 별로 독립적으로 분할이 가능하기 때문에 YOLO[2]와 같은 사각 영역 기반 검출기 대비 차종이나 도로 표지판, 보행자 인식 등에 유용하게 적용 가능하다. Mask R-CNN은 다양한 도로 환경에서 기존 심층신경망 방법 대비 높은 검출 성능을 보여주지만, 2단계 구조로 인한 많은 수의 신경망 매개변수 및 이에 상응하는 GPU 메모리가 요구된다. 이는 Mask R-CNN의 임베디드 환경에서 동작을 어렵게 하며 따라서 Mask R-CNN은 실제 자율주행을 위한 자동차에 서버와의 통신 없이 탑재되기 어렵다. 또한, 영역 제안 신경망을 통해 추출된 잠재 특징이 서로 다른 세 가지 작업에 동일하게 사용됨으로써 각 작업별 특징의 중요도를 고려하지 않는 문제점이 있다.

본 논문에서는 이러한 한계점을 극복하기 위하여 기존 모델의 검출 성능은 유지하면서 심층신경망 모델의 매개변수 수를 대폭 줄인 경량화 모델을 제안한다. 제안하는 방법은 특징 압축을 위해 경량화에 뛰어난 성능을 보이는 EfficientNet[3] 구조를 Backbone 신경망으로 적용하였으며, Mask R-CNN과 달리 다중 스케일 정보를 효과적으로 사용하여 특징을 추출하기 위해 BiFPN(Bi-directional FPN)[4]을 적용하였다. 또한, 주의 모듈(Attention Module)[5]을 각 작업별 가지 신경망(Branch Network)에 삽입함으로써 도출된 특징의 중요도가 각 작업의 목적에 맞게 재조정(Re-calibration)되도록 한다. 이를 통해 총 매개변수 수를 효과적으로 감소시킬 수 있으며, 검출 성능 또한 유지할 수 있다. Benchmark 데이터셋에 대한 실험을 통해 제안하는 경량화 방법이 총 매개변수 수를 Mask R-CNN 대비 약 절반으로 감소시킬 수 있으며 주의 모듈을 이용한 특징 재조정을 기반으로 검출 성능 또한 효과적으로 유지 할 수 있음을 확인할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 제안하는 경량화 심층신경망 구조 및 주의 모듈 기반의 특징 결합 구조에 대해 자세히 설명하며, 3장에서는 다양한 실험을 통해 제안하는 방법이 기존방법 대비 도로 객체 검출 성능이 개선됨을 검증한다. 마지막으로 4장에서는 본 논문의 결론을 서술한다.

a) 건국대학교 전기전자공학부(Department of Electrical and Electronics Engineering, Konkuk University)
b) 한국과학기술정보연구원 연구데이터공유센터(Research Data Sharing Center, Korea Institute of Science and Technology Information)
‡ Corresponding Author : 김원준(Wonjun Kim)
E-mail: wonjkim@konkuk.ac.kr
Tel: +82-2-450-3396
ORCID: https://orcid.org/0000-0001-5121-5931
※ 본 연구는 한국과학기술정보연구원(KISTI) “연구데이터 공유 확산체제 구축(K-20-L01-C04-S01)” 과제의 위탁연구로 수행한 것입니다.
※ This work was supported by a Research and Development project, “Enabling a System for Sharing and Disseminating Research Data,” of Korea Institute of Science and Technology Information (KISTI), South Korea, under Grant K-20-L01-C04-S01.
· Manuscript received June 22, 2020; Revised October 7, 2020; Accepted October 7, 2020.

Ⅱ. 제안하는 방법

제안하는 방법은 Mask R-CNN 구조를 효율적으로 개량하여 객체 검출 성능을 유지하면서 신경망 모델의 매개변수 수를 대폭 줄이는 데 집중한다. 구체적으로, 적은 수의 매개변수를 이용하여 효과적으로 영상 특징을 추출할 수 있는 EfficientNet 구조[3]를 Backbone으로 적용하고, 다양한 스케일의 정보를 효과적으로 결합하기 위해 BiFPN 구조[4]를 함께 이용 하였다. 또한, 객체 종류 인식, 객체 위치 예측 및 영역 분할의 중요도에 따라 추출된 특징 값을 재조정하기 위해 각 작업을 위한 가지(Branch) 신경망의 앞단에 주의 모듈(Attention Module)을 삽입한다. 본 장에서는 먼저 제안하는 신경망의 전체적인 경량화 구조에 대해 소개한다. 이어서 주의 모듈 기반의 특징 결합 방식에 대해 자세히 설명한 후 마지막으로 제안하는 구조를 기반으로 객체 검출 학습에 사용된 손실 함수에 대해 설명한다.

1. Mask R-CNN 경량화 모델의 구조

제안하는 방법의 전체적인 구조는 그림 1과 같다. EfficientNet[3] 구조를 기반으로 영상 특징을 다중 스케일에서 추출하는 Backbone으로 사용하였다. EfficientNet은 모델의 크기와 연산량을 결정하는 세 가지 요소(즉, 해상도, 깊이, 너비)를 동시에 고려하는 복합 스케일링(Compound Scaling) 기법을 사용하여 효과적으로 특징을 압축한다. 해상도(Resolution)는 입력 영상의 크기를 의미하며, 깊이(Depth)는 신경망 계층의 개수, 너비(Width)는 필터, 즉 채널의 개수를 의미한다. 기존의 Mask R-CNN에서 사용하는 ResNet-101 구조는 영상 특징이 압축되면서 채널의 개수가 2,048까지 증가하는 반면에, 본 논문에서는 채택한 EfficientNet-B4 구조를 기반으로 영상 특징 압축을 수행할 시 224채널만을 사용하게 되며 이에 따라 모델의 학습 매개변수는 2,300만개 가량 줄어들게 된다. Backbone 구조의 각 스케일마다 특징맵(Feature Map)을 추출하여 다섯 개의 특징맵 $P_k\,(k=3,4,5,6,7)$을 얻는다. 이러한 특징 추출 과정은 다음과 같이 표현될 수 있다.

$$P_k = B_k(I),\quad k=3,4,5,6,7, \tag{1}$$

여기서 I는 입력 영상이며, $B_k(I)$는 Backbone 신경망을 통해 획득한 입력 영상의 $1/2^k$ 해상도에 해당하는 특징맵을 의미한다 (그림 1 참조). 또한, 스케일 간 정보를 조밀하게 결합하여 사용하기 위해 기존의 FPN 구조를 BiFPN[4] 구조로 교체하였다. 높은 수준에서 낮은 수준의 방향으로만 연결선이 존재하는 하향식 방식의 FPN 구조와 달리, BiFPN은 상향식 방식의 경로가 추가되어 모든 스케일에서 추출된 특징들을 풍성하게 결합한다. 제안하는 신경망 구조에서는 BiFPN을 여러 번 반복하여 객체의 다양한 변이

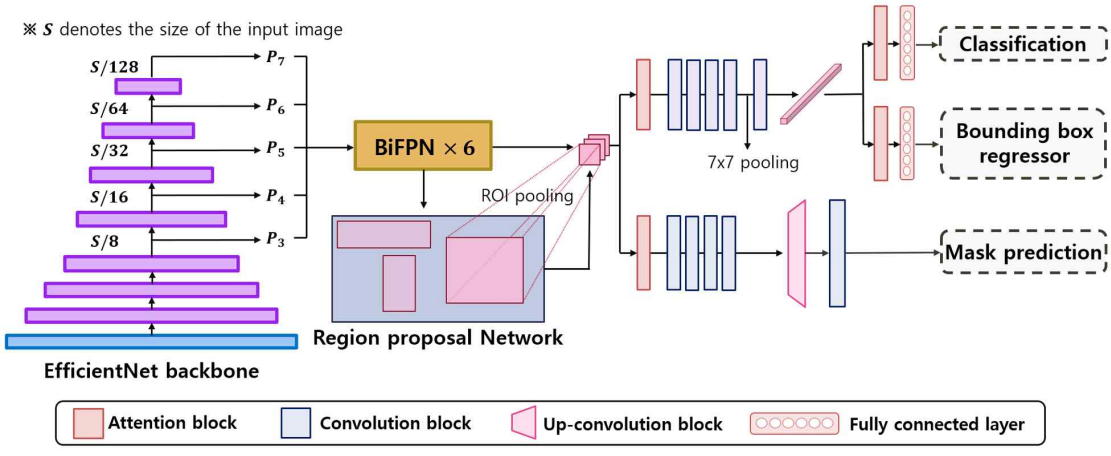


그림 1. 객체 검출을 위한 제안하는 신경망의 전체적인 구조
Fig. 1. Overall architecture of the proposed network for object detection

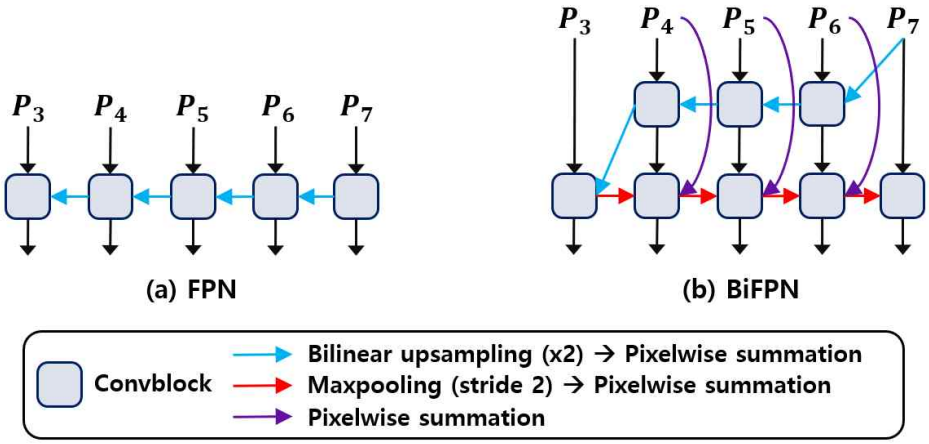


그림 2. 본 논문에서 사용된 BiFPN의 상세 구조
Fig. 2. Details of the BiFPN structure for the proposed network architecture

를 효과적으로 고려할 수 있는 고차원 잠재 특징이 추출될 수 있도록 하였다. BiFPN의 자세한 구조는 그림 2에 나타내었다.

영역 제안 신경망(Region Proposal Network)은 BiFPN을 통하여 추출된 잠재 특징 전체 영역에서 앵커 박스(Anchor Box)를 이용하여 객체 후보 영역들을 생성하고 관심 영역 풀링(ROI Pooling)을 통해 일정한 해상도의 지역적 특징맵을 생성한다. 본 논문에서는 관심 영역 풀링 기법으로 Mask R-CNN[1]에서 사용되었던 ROI-Align을 채택하였다. 해당 지역적 특징맵은 주의 모듈을 통해 세 개의 가지(Branch) 신경망을 거쳐 객체 인식과 객체 위치 예측 및 영역 분할에 사용된다 (그림 1 참조). 가지 신경망의 구조를 결정할 때에도 앞서 설명한 복합 스케일링 기법을 적용하여 최적의 합성곱 신경망 계층의 수와 채널의 수를 선택하였다. 이를 통해 기존의 Mask R-CNN에 사용된 작업별 신경망의 매개변수보다 1,000만개 가량 줄어든 효율적인 신경망 구조를 구축하였다.

2. 주의 모듈을 이용한 특징 결합 재조정

기존 Mask R-CNN에서는 객체 인식(Classification), 객체 위치 예측(Bounding Box Regression) 및 영역 분할(Segmentation)의 세 가지 작업을 위해 동일한 지역적 특징 사용하기 때문에 각 작업별 중요도가 고려되지 않는다. 제안하는 신경망 구조에서는 각 작업을 위한 가지 신경망마다 주의 모듈(Attention Module)을 삽입함으로써 객체 영역 내 지역적 특징 값이 해당 작업의 목적에 맞게 재조정 될 수 있도록 하였다. 본 논문에서는 지역적 특징의 채널 별

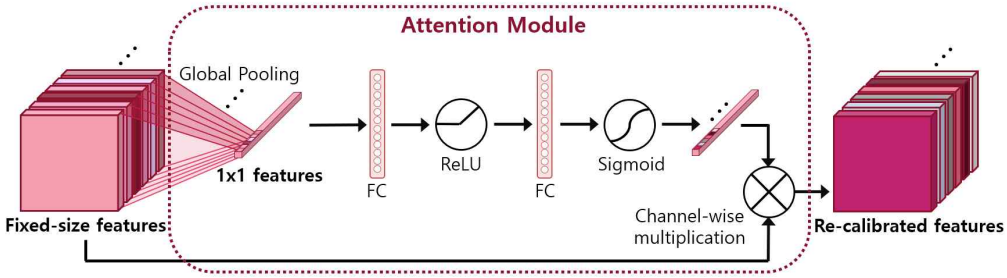


그림 3. 본 논문에서 사용된 주의 모듈의 상세 구조
Fig. 3. Details of the attention-module for the proposed network architecture

재조정을 위해 SE Block[5]을 사용하였다. 구체적으로, SE Block 기반의 주의 모듈은 입력된 지역적 특징을 전역 평균 풀링(Global Average Pooling)을 통해 채널 단위로 1×1 값(즉, Scalar)이 되도록 특징 정보를 압축한다. 그 후, 그림 3에서 볼 수 있듯이 완전 연결 계층(Fully Connected Layer), ReLU(Rectified Linear Unit), 완전 연결 계층, 그리고 시그모이드(Sigmoid) 함수를 거쳐 압축된 특징의 중요도를 채널 단위로 계산한다. 계산된 중요도는 가중치로써 입력된 지역적 특징의 각 채널에 곱해져 특징 값을 새롭게 재조정하는 역할을 한다. 따라서 본 논문에서는 관심 영역 풀링이 적용된 잠재 특징을 먼저 단순 결합한 이후, 그림 1에서 볼 수 있듯이 작업별 가지 신경망마다 각각 주의 모듈을 적용하여 해당 작업에 적합한 특징의 채널별 중요도를 학습을 통해 도출하였다. 도출된 채널별 중요도를 이용하여 작업별 가지 신경망에 입력되기 전에 잠재 특징을 재조정해주었다. 3장의 실험 결과 및 분석을 통해 각 작업의 목적에 맞게 주의 모듈을 적용하는 것이 세 가지 작업을 효과적으로 수행하는 데 도움이 되며 검출 성능 향상에 기여함을 보인다.

3. 객체 검출을 위한 손실 함수 설계

본 논문에서는 주의 모듈이 적용된 EfficientNet 구조와 기존 Mask R-CNN 구조의 차이에 의한 매개변수 감소에 초점을 맞추었으며, 손실 함수는 기존 Mask R-CNN에서 사용하던 방식을 그대로 채택하였다. 간단히 살펴보면, 신경망의 최종 예측 값에 적용되는 손실 함수 L은 객체 인식에 이용되는 L_{ds}, 객체 위치 예측에 이용되는 L_{reg} 및 객체별 분할에 이용되는 L_{mask}의 합으로 정의되며 다음과 같이 표현될 수 있다.

$$L = L_{ds} + L_{reg} + L_{mask}, \tag{2}$$

여기서 L_{ds}는 객체의 실제 클래스 값과 예측 값의 차이를 의미하며, 교차 엔트로피(Cross Entropy)를 이용하여 다음과 같이 계산된다.

$$L_{ds} = -\sum_{i} u_i \log p_i, \tag{3}$$

여기서 u_i는 관심 영역 풀링을 통해 추출된 관심 영역들 중 i번째 영역의 실제 클래스 값이며, p_i는 신경망을 통해 예측된 클래스 값이다. 다음으로 추출된 관심 영역들 중 클래스가 배경인 경우를 제외한 영역들을 사용해 L_{reg}를 계산하며 아래와 같이 표현된다.

$$L_{reg} = \sum_{i} \sum_{k\in[x,y,w,h]} [u_i \geq 1] \mathrm{smooth}_{L1}(t_i^{u_i}(k) - v_i(k)), \tag{4}$$

여기서 v_i는 클래스 u_i에 해당하는 실제 객체 상자의 위치 정보이며, $t_i^{u_i}$는 예측된 객체 상자의 위치 정보에 해당한다. $[x,y,w,h]$은 각각 객체 상자 좌측 상단에 해당하는 x좌표와 y좌표, 그리고 너비 및 높이 값을 나타낸다. $[u_i \geq 1]$은 클래스 u_i가 배경이 아닐 때는 1이 할당되고 배경일 때는 0이 할당되어 손실 함수 계산 과정에서 배경에 해당하는 관심 영역을 제외한다. 또한, smooth_{L1}은 변형된 $L1$ 손실 함수로서 $L2$ 손실 함수에 비해 이상치(Outlier) 값에 덜 민감하며, 아래와 같이 계산된다.

$$\mathrm{smooth}_{L1}(x) = \begin{cases} 0.5x^2 & \mathrm{if}|x| < 1, \\ |x| - 0.5 & \mathrm{otherwise}. \end{cases} \tag{5}$$

마지막으로 L_{mask}는 각 객체 상자 내에서 예측된 이진 마스크 결과 영상과 실제 마스크 영상과의 비교를 통해 계산된다. 이를 위해 이진 교차 엔트로피 함수가 사용되었으며, 아래의 식을 이용하여 계산된다.

$$L_{mask} = -\sum_{i}\sum_{m}^{H}\sum_{n}^{W} q_i(m,n)\log p_i(m,n), \tag{6}$$

여기서 H와 W는 각각 마스크 영상의 높이와 너비 값을 나타낸다. p_i는 예측된 마스크 영상의 픽셀 단위 확률 값이며, q_i는 실제 마스크 영상의 픽셀 단위 클래스 값으로서 해당 픽셀이 객체이면 1, 객체가 아니면 0으로 할당된다.

Ⅲ. 실험 결과 및 분석

본 논문에서는 제안하는 방법의 성능 평가를 위해 두 개

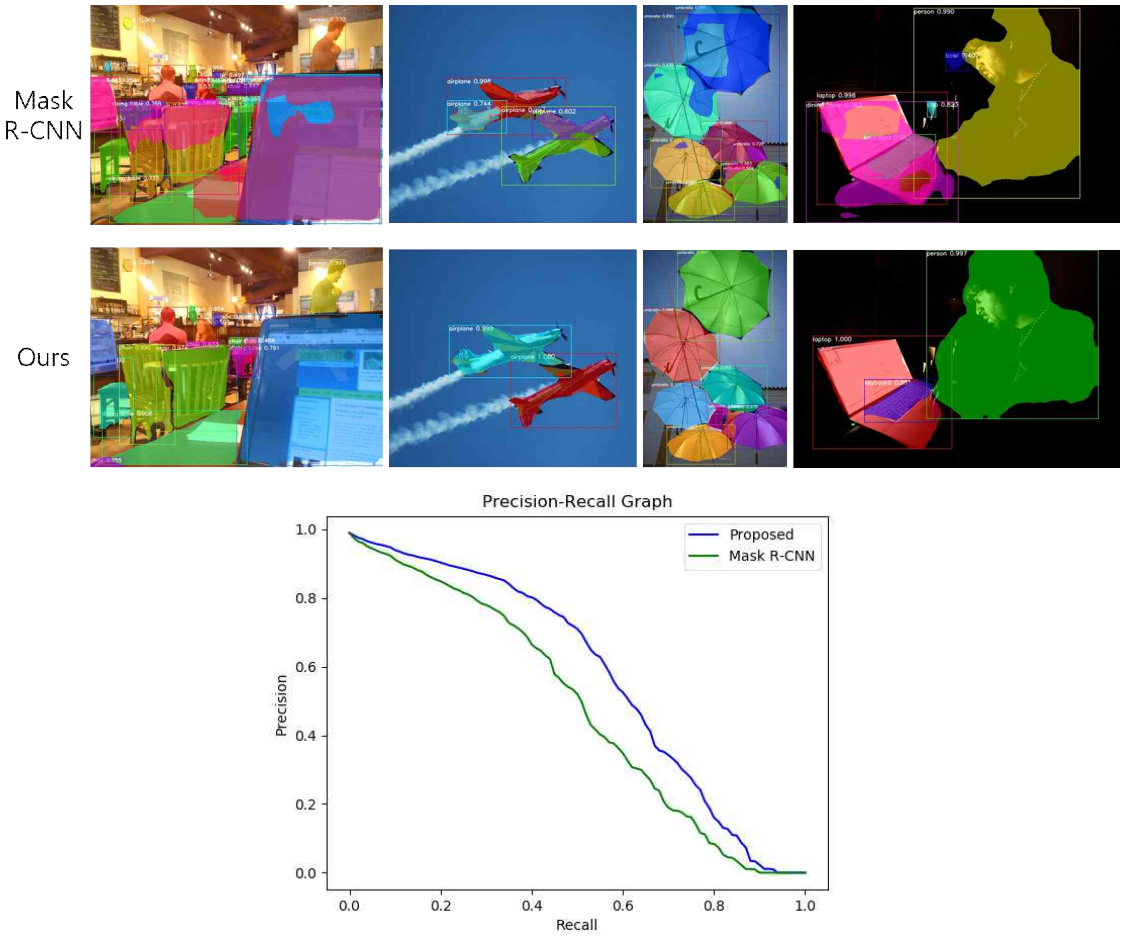


그림 4. MS COCO[6] test-dev 데이터셋에서의 객체 검출 결과
Fig. 4. Results of the object detection on the MS COCO[6] test-dev dataset

의 벤치마크 데이터셋을 사용하였다. 첫 번째로는 객체 검출 분야에서 일반적으로 사용되는 MS COCO 2017 검출 데이터셋[6]를 사용하였다. 학습을 위해 총 12만장의 이미지가 사용되며, 5,000장의 이미지가 학습 중 성능 검증(Validation)을 위해 사용되었다. MS COCO에서 제공하는 총 4,000장의 이미지로 구성된 test-dev 데이터셋을 사용하여 최종 성능 평가를 수행하였다. MS COCO 데이터셋 내의 객체들은 기본적으로 80개의 클래스 정보를 가지고 있으며, 사람, 동물, 탈것, 음식, 사물 등 다양한 객체 종류를 포함하고 있다. 두 번째로는 국내 도로 환경에서 제안하는 모델의 성능을 평가하기 위해 한국과학기술정보연구원에서 제공하는 도로 주행 동영상에서 총 7종의 객체를 추출하여 KISTI 데이터셋으로 구축하고 성능 평가에 활용하였다. 실험에 사용된 영상들은 1920×1440 화소의 크기를 가지며 대로변, 시장, 번화가 등 다양한 환경에서 촬영된 영상을 포함하고 있다. 도로 주행 영상에서 총 1,300장의 이미지를 추출하였으며, 이 중 1,200장의 이미지가 학습을 위해 사용하였고 100장의 이미지는 성능 평가를 위해 사용하였다. 구축된 데이터셋은 도로 환경에서 가장 빈번하게 검출되는 7종 객체, 즉, 사람, 차, 버스, 트럭, 오토바이, 공사 표지판, 포트홀을 포함하고 있다.

제안하는 방법은 파이토치(Pytorch)[7] 프레임워크에 기반하여 구현되었다. Backbone 신경망은 EfficientNet-B4[3]을 사용하였고, 신경망의 학습 매개변수들은 ILSVRC[8] 데이터셋을 이용하여 미리 학습된 모델의 매개변수 값으로 초기화되었다. 영역 제안 신경망과 BiFPN 및 다중 작업(즉,

그림 5. 한국과학기술정보연구원의 도로 주행 영상을 기반으로 구축된 데이터셋에서의 객체 검출 결과
Fig. 5. Results of the object detection on the KISTI dataset

객체 인식, 객체 위치 예측, 영역 분할)을 수행하는 신경망의 매개변수들은 [9]에 소개된 방법을 이용하여 초기화되었다. 다중 작업 신경망의 정규화 계층으로는 배치(Batch) 크기에 무관하다고 알려진 그룹 정규화(Group Normalization)[10] 기법을 사용하였고, 배치 크기는 1로 설정되었다. 본 논문에서 손실 함수를 최적화하기 위한 알고리즘으로는 AdamW[11]을 사용하였고, 파워(Power)와 가속도(Momentum) 값은 각각 0.9와 0.999로 설정하였다. 가중치 감쇠(Weight Decay) 값은 Backbone 신경망에서는 0.0005로 설정하였으며 그 외의 신경망에서는 0으로 설정하였다. 학습 속도(Learning Rate)는 10^{-4}부터 시작하여 다항 감쇠(Polynomial Decay) 스케줄링을 통해 학습 데이터셋을 200회 반복할 동안 10^{-5}까지 감소한다. 1회 학습을 반복할 동안 2,000장의 이미지가 학습 데이터셋로부터 임의의 확률로 배치 크기(즉, 1장) 단위로 추출된다. 학습과 성능 평가에는 Geforce GTX TITAN X 1개가 이용되었다. 모든 학습 이미지는 신경망에 입력되기 전 원본 이미지 크기의 비율을 유지하면서 최소 800픽셀, 최대 1024픽셀의 크기로 재조정되고, 0.5의 확률로 좌우 반전된다.

제안하는 방법의 효율성을 검증하기 위해 본 논문에서는 mAP(mean Average Precision)값을 활용하여 정량적 성능 비교를 수행하였다. 여기서 mAP는 값이 클수록 좋은 성능을 의미한다. MS COCO 데이터셋과 한국과학기술정보연

표 2. 한국과학기술정보연구원으로부터 구축된 KISTI 데이터셋에서의 정량적 평가 비교
Table 2. Quantitative evaluations on the KISTI dataset

Methods	Params	AP	AP_{50}	AP_{75}
Mask R-CNN[1]	69.59M	63.8	73.5	65.3
Proposed method	32.71M	69.6	88.7	74.7

표 1. MS COCO[6] test-dev 데이터셋에서의 정량적 평가 비교
Table 1. Quantitative evaluations on the MS COCO[6] test-dev dataset

Method	Params	Speed(fps)	AP	AP_{50}	AP_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
FCIS[12]	63.32M	-	-	-	-	33.6	54.5	-
Mask R-CNN[1]	69.59M	5.1	38.2	60.3	41.7	35.7	58.0	37.8
Proposed method	32.71M	7.2	40.4	62.7	44.5	37.1	58.6	38.2

구원의 도로 주행 영상으로부터 구축한 KISTI 데이터셋에서의 성능 분석 결과는 표 1과 2에 각각 정리하여 나타내었다. AP_{50}은 정답이라고 판단되는 IoU(Intersection of Union) 값의 기준이 0.5이상일 때 mAP를 의미하며, AP_{75}는 정답에 해당하는 IoU 값의 기준이 0.75이상일 때 mAP를 의미한다. AP는 IoU의 기준을 0.5부터 시작하여 0.05씩 늘려가면서 0.95까지 각각의 기준에서 mAP를 측정한 수치의 평균 mAP를 의미한다. AP^{mask}는 객체 영역 분할(Instance Segmentation) 결과에 대한 평균 mAP를 의미한다. 표 1과 2 모두에서 제안하는 방법이 더 적은 매개변수로 기존 방법들의 성능을 상회함을 볼 수 있다. 특히, 도로 주행 영상 내 객체 검출을 위한 KISTI 데이터셋에서 큰 성능 향상을 보이면서 제안하는 방법이 자율 주행 시스템을 위한 효율적인 장면 인식 솔루션으로 사용될 수 있음을 확인할 수 있다. 정성적인 비교를 위해 그림 4에 객체 검출 결과의 예를 나타내었다. 그림에서 볼 수 있듯이, 복잡한 배경 영역에서도 제안하는 방법이 객체의 경계를 정밀하게 분할하고 있다. 특히, 두 번째 예제 이미지에 대해서 제안하는 방법은 기존 방법에 비해 비행기의 영역을 온전히 검출하였다. 한국과학기술정보연구원의 도로 주행 영상을 기반으로 구축 된 KISTI 데이터셋에서의 객체 검출결과는 그림 5에 나타내었다. 제안하는 방법이 낮은 조도 환경과 객체가 겹쳐있는 상황(예를 들어, 위에서 첫 번째 예시와 아래에서 첫 번째 예시)에서도 안정적으로 객체를 검출할 수 있음을 볼 수 있다. 또한, 제안하는 방법에서 사용된 BiFPN 구조는 FPN과 비교하여 1.7M의 매개변수 증가 비용이 있지만 mAP 성능을 1.1만큼 크게 향상시켰다. 최종적으로 주의 모듈을 이용하여 객체 영역 내 특징을 적응적으로 재조정하는 방법의 우수성을 보이기 위하여 주의 모듈을 적용한 신경망 구조와 주의 모듈을 적용하지 않은 신경망 구조에 대하여 각각 성능 평가를 수행하였으며, 그 결과를 표 3에 나타내었다. 표의 결과에서 알 수 있듯이 모든 데이터셋에서 주의 모듈을 적용하였을 때가 그렇지 않을 때보다 적은 매개변수 증가 비용으로 큰 검출 성능 향상을 달성함을 볼 수 있다. 따라서, 제안하는 주의 모듈 기반 Mask R-CNN 경량화 모델이 도로 주행 환경 내 주요 객체를 검출하기 위한 효율적 솔루션이 될 수 있음을 확인할 수 있다.

Ⅳ. 결 론

본 논문에서는 주의 모듈 기반의 Mask R-CNN 경량화 모델을 제안하였다. 제안하는 방법은 효율적인 특징 압축을 위해 Backbone 신경망 구조를 수정하고 기존의 단순 FPN 구조를 다중 스케일의 특징 추출에 최적화된 BiFPN으로 대체하였다. 다중 작업 신경망에서도 복합 스케일링 기법을 통해 최적의 합성곱 계층과 채널 수를 찾아 제안하는 신경망 구조에 적용하였다. 또한, 주의 모듈을 적용하여 다중 작업 신경망에 입력되는 객체 영역 특징을 각 작업의 목적에 맞게 적응적으로 재조정시켜 효과적인 객체 검출 학습을 가능하게 하였다. 다양한 실험을 통해 제안하는 방법이 기존 방법 대비 신경망 모델의 매개변수를 대폭 감소시키고 검출 성능 또한 향상시켰음을 보였다.

참 고 문 헌 (References)

[1] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” in *Proc. IEEE International Conference on Computer Vision(ICCV)*, pp. 2980 - 2988, Oct. 2017.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look

표 3. 주의 모듈 적용 여부에 따른 제안하는 신경망 구조의 정량적 성능 비교
Table 3. Performance variation of the proposed method according to the attention module

Methods	Dataset	Params	AP	AP_{50}	AP_{75}
Without Attention Module	MS COCO	31.53M	38.1	58.4	42.2
With Attention Module	MS COCO	32.71M	40.4	62.7	44.5
Without Attention Module	KISTI	31.53M	67.3	84.9	70.5
With Attention Module	KISTI	32.71M	69.6	88.7	74.7

once: Unified real-time object detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 779-788, Jun. 2016.

[3] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th International Conference on Machine Learning*, pp. 6105 – 6114, Jun. 2019.

[4] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," 2019, *arXiv:1911.09070.* [Online]. Available: http://arxiv.org/abs/1911.09070

[5] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, pp. 7132-7141, Jun. 2018.

[6] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and L. Zitnick. "Microsoft coco: Common objects in context." in *Proc. European Conference on Computer Vision (ECCV)*, pp. 740-755, 2014.

[7] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. Devito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, "Automatic differentiation in pytorch". in *Proc. Conference and Workshop on Neural Information Processing Systems (NIPS)*, pp. 1-4, 2017.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211 – 252, Dec. 2015.

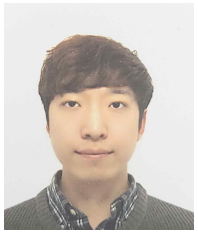
[9] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 1026 – 1034, Dec. 2015.

[10] Y. Wu and K. He, "Group normalization," in P*roc. European Conference on Computer Vision (ECCV)*, pp. 3 – 19, Sep. 2018.

[11] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv:1711.05101*. [Online]. Available: https://arxiv.org/abs/1711.05101, 2017.

[12] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. "Fully convolutional instance-aware semantic segmentation," in *Proc. IEEE International Conference on Computer Vision Pattern Recognition (CVPR)*, pp. 4438-4446, Jul. 2017.

저 자 소 개



송 민 수

- 2020년 2월 : 건국대학교 학사
- 2020년 3월 ~ 현재 : 건국대학교 전기전자공학부 석사과정
- ORCID : https://orcid.org/0000-0003-3823-4913
- 주관심분야 : 컴퓨터 비전, 객체 검출, 기계학습, 패턴 인식

김 원 준

- 2012년 8월 : 한국과학기술원(KAIST) 박사
- 2012년 9월 ~ 2016년 2월 : 삼성종합기술원 전문연구원
- 2016년 3월 ~ 2020년 2월 : 건국대학교 전기전자공학부 조교수
- 2020년 3월 ~ 현재 : 건국대학교 전기전자공학부 부교수
- ORCID : https://orcid.org/0000-0001-5121-5931
- 주관심분야 : 영상이해, 컴퓨터 비전, 기계학습, 패턴 인식

저 자 소 개

장 래 영

- 2018년 8월 : 한남대학교 컴퓨터공학과 박사
- 2019년 9월 ~ 현재 : 한국과학기술정보연구원 연구데이터공유센터 박사후연구원
- 주관심분야 : DevOps, Docker, Kubernetes, 인공지능

이 용

- 2003년 : 일본교토대학교 사회정보학과 박사
- 2011년 ~ 2013년 : 일본NICT연구소 연구원
- 2013년 9월 ~ 현재 : 한국과학기술정보연구원 연구데이터공유센터 책임연구원
- 주관심분야 : 공간데이터, 사물인터넷, 스마트시티, 인공지능

박 민 우

- 2004년 : 충남대학교 컴퓨터과학과 석사
- 1996년 8월 ~ 현재: 한국과학기술정보연구원 연구데이터공유센터 책임연구원
- 주관심분야 : 시스템아키텍쳐, 정보보안, 인공지능

이 상 환

- 2018년 : 서울시립대학교 컴퓨터공학과 박사
- 1995년 4월 ~ 현재 : 한국과학기술정보연구원 연구데이터공유센터 책임연구원
- 주관심분야 : 빅데이터 분석, 데이터 생태계, 인공지능

최 명 석

- 2005년 8월 : 한국과학기술원(KAIST) 전산학과 박사
- 2005년 6월 ~ 현재 : 한국과학기술정보연구원 연구데이터공유센터 선임연구원
- 주관심분야 : 오픈사이언스, 연구데이터 관리, 인공지능