JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 24, No. 12: 1581~1587, Dec. 2020

주의 집중 기법을 활용한 객체 검출 모델

김근식[1] · 배정수[2] · 차의영[3*]

Object Detection Model Using Attention Mechanism

Geun-Sik Kim[1] · Jung-Soo Bae[2] · Eui-Young Cha[3*]

[1]Graduate Student, Department of Information Convergence Engineering, Pusan National University, Busan, 46241 Korea
[2]Assistant Professor, College of Software Convergence, Dongseo University, Busan, 47011 Korea
[3*]Professor, Department of Computer Engineering, Pusan National University, Busan, 46241 Korea

요 약

기계 학습 분야에 합성 곱 신경망이 대두되면서 이미지 처리 문제를 해결하는 모델은 비약적인 발전을 맞이했다. 하지만 그만큼 요구되는 컴퓨팅 자원 또한 상승하여 일반적인 환경에서 이를 학습해보기는 쉽지 않은 일이다. 주의 집중 기법은 본래 순환 신경망의 기울기 소실 문제를 방지하기 위해 제안된 기법이지만, 이는 합성 곱 신경망의 학습에도 유리한 방향으로 활용될 수 있다. 본 논문에서는 합성 곱 신경망에 주의 집중 기법을 적용하고, 이때의 학습 시간과 성능 차이 비교를 통해 제안하는 방법의 우수성을 입증한다. 제안하는 모델은 YOLO를 기반으로 한 객체 검출에서 주의 집중 기법을 적용하지 않은 모델에 비해 학습 시간, 성능 모두 우수한 것으로 나타났으며, 특히 학습 시간을 현저히 낮출 수 있음을 실험적으로 증명하였다. 또한, 이를 통해 일반 사용자의 기계 학습에 대한 접근성 증대가 기대된다.

ABSTRACT

With the emergence of convolutional neural network in the field of machine learning, the model for solving image processing problems has seen rapid development. However, the computing resources required are also rising, making it difficult to learn from a typical environment. Attention mechanism is originally proposed to prevent the gradient vanishing problem of the recurrent neural network, but this can also be used in a direction favorable to learning of the convolutional neural network. In this paper, attention mechanism is applied to convolutional neural network, and the excellence of the proposed method is demonstrated through the comparison of learning time and performance difference at this time. The proposed model showed that both learning time and performance were superior in object detection based on YOLO compared to models without attention mechanism, and experimentally demonstrated that learning time could be significantly reduced. In addition, this is expected to increase accessibility to machine learning by end users.

키워드 : 기계 학습, 객체 검출, 주의 집중 기법, 합성 곱 신경망

Keywords : Machine learning, Object detection, Attention mechanism, CNN(Convolutional neural network)

Received 8 September 2020, **Revised** 14 September 2020, **Accepted** 29 September 2020

***Corresponding Author** Eui-Young Cha(E-mail:eyecha@pusan.ac.kr, Tel:+82-51-510-2219)
Professor, Department of Computer Engineering, Pusan National University, Busan, 46241 Korea

Open Access http://doi.org/10.6109/jkiice.2020.24.12.1581 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

기계 학습(Machine learning) 분야에 합성 곱 신경망(Convolutional neural network)이 대두되면서 이미지 처리 문제는 큰 변화를 맞이했다. 기존의 다층 퍼셉트론(Multi layer perceptron)으로도 이미지 처리에 대한 학습은 가능하지만, 규모가 커질수록 학습해야 하는 데이터의 크기와 학습 시간이 매우 커지게 되며, 이미지의 위치, 각도, 크기 변화에도 취약해진다. 합성 곱 신경망은 이러한 취약점을 보완하면서 학습 시간 또한 크게 줄일 수 있었기에 이미지 처리 문제에서 주류로 자리 잡을 수 있었다.

이러한 기술의 발전에도 불구하고 이미지 처리 연산에는 많은 컴퓨팅 자원과 시간이 소모된다. 때문에 일반적인 환경에서 특히, 사용자 입장에서 이를 직접적으로 시도해보는 것은 부담스러운 일일 것이다.

주의 집중 기법(Attention mechanism)은 본래 자연어 처리(Natural language processing)와 기계 번역(Machine translation)을 위한 순환 신경망 기반(Recurrent neural network)의 Seq2Seq 모델의 문제점을 보완하기 위해 제안되었다. 순환 신경망에는 기울기 소실(Gradient vanishing)이라는 문제가 존재하는데, 이로 인해 망이 깊어질수록 역 전파 과정에서 입력 층 근처의 기울기가 점차 작아지는 현상이 발생할 수 있다. 결국, 이는 가중치 갱신을 방해해 최적의 모델을 찾을 수 없도록 한다.

이러한 기울기 소실 문제를 해결하기 위해 주의 집중 기법은 전체 입력을 동일한 비율로 참고하는 것이 아니라, 특정 부분에 가중치를 주어 학습을 진행함으로써 해당 문제를 해결함과 동시에 모델의 성능 또한 비약적으로 향상할 수 있었다.

본 논문에서는 합성 곱 신경망에 주의 집중 기법을 적용하여 이미지 처리 문제에서의 학습 시간을 줄이고, 좀 더 최적화된 모델을 학습할 수 있도록 하는 방법을 제안한다. 백본 모델(Backbone model)로 YOLO(You only look once)[1]의 Darknet을 사용하였으며, 병목 층(Bottleneck layer)에 주의 집중 기법을 적용하였다. 그리고 백본 모델을 단독으로 사용했을 때와 주의 집중 기법을 적용하였을 때의 학습 시간과 성능 차이 비교를 통해 제안하는 방법의 우수성을 입증한다.

Ⅱ. 관련 연구

2.1. 주의 집중 기법

인간은 눈으로 직접 어떠한 풍경을 볼 때, 시야에 들어오는 모든 부분을 주목해서 보지는 않는다. 즉, 이미지의 특정 부분에 대해서는 고해상도로 집중하는 반면, 그 외의 주변 부분은 저해상도로 인식하고, 이후에 초점 영역을 조정하여 전체 이미지에 대해 추론을 한다.

주의 집중 기법이란, 이러한 인간의 추론 과정을 인공 신경망에 적용하여 예측 성능을 높이고, 인간이 신경망의 동작을 좀 더 직관적으로 이해하는 것이 가능하도록 하려는 시도 중의 하나이다.

합성 곱 신경망에서 주의 집중이라는 개념은 시각적 질의응답(Visual question answering)과 이미지 캡셔닝(Image captioning)에서 주로 사용되어왔다. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention"[2]에서는 입력 이미지의 어떤 부분에 주목하여 문장을 생성하였는지 효과적으로 시각화할 수 있었고, "Dual Attention Networks for Multimodal Reasoning and Matching"[3]에서는 이미지에서 질문으로, 질문에서 이미지로 이어지는 양방향 주의 집중 기법을 활용하여 시각적 질의응답의 성능을 향상시켰다.

위와 같은 예는 주의 집중 기법을 활용한 신경망에서 입력에 따라 동적 특징 선택(Dynamic feature selection)이 일어난다는 것을 보여준다. 특징 선택이란, 모든 특징 중에서 필요한 부분만을 선택하여 간결한 특징 집합을 만드는 것을 말한다. 이미지를 예로 들면, 전체 픽셀에서 결과를 예측하는데 큰 영향을 미치지 않는다고 생각되는 부분을 배제하거나, 비중을 적게 두어 학습을 계속하는 것이다.

이러한 과정은 일반적인 이미지 인식이나 객체 검출 문제에도 동일하게 적용할 수 있다. 여기에서도 입력에 따라 주목해야 할 부분이 각자 다를 것이고, 학습을 통해 성능 향상을 기대할 수 있을 것이다.

망의 구조적 관점에서의 주의 집중 기법은 비교적 최근부터 연구되기 시작했다. RAN(Residual attention network)[4]은 주의 집중이 적용된 신경망을 모듈화하고, 여러 층으로 쌓아 올려 이미지 인식에서 성능 향상을 보였지만 복잡한 구조로 인해 연산량이 많아지는 문제점이 있었다. 이후에 등장한 SENet(Squeeze and excitation networks)[5]은 문자 그대로 압축과 재조정을 통해 이미

지의 각 채널에서 중요한 특징만을 추출하고, 이를 가중치로 두어 학습함으로써 모델 오류를 상당히 떨어뜨릴 수 있었다. 여기에 CBAM(Convolutional block attention module)[6]은 채널 정보뿐만 아니라 공간 정보에 대한 주의 집중 연산까지 추가하여 성능 향상 및 학습 시간 단축을 보였으며, ECA-Net(Efficient channel attention networks)[7]은 연산의 차원 축소(Dimensionality reduction)를 배제하여 채널에 대한 주의 집중 연산을 개선하였다.

대부분의 최신 연구는 주의 집중 연산을 모듈화하여 일반적인 신경망에 끼워 넣음으로써 오버헤드와 모델 복잡도를 최소화하면서도 성능 향상을 꾀하는 방향으로 진행되고 있다.

2.2. IoU(Intersection over union)

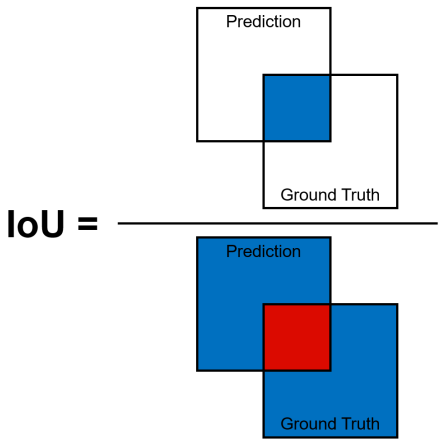

Fig. 1 Intersection over union.

두 영역이 얼마나 겹쳐 있는지 판단하는 척도로 객체 검출에서 예측된 경계 상자의 정확도를 평가하는 지표 중의 하나이며, 그림 1과 같이 예측된 경계 상자와 실제 영역(Ground truth)이 겹치는 영역을 전체 영역으로 나누고, 이 값을 정확도로 간주한다. 일반적으로 값이 0.5 이상이면 제대로 검출되었다고 판단하고, 그렇지 않다면 잘못 검출되었다고 판단한다. 이 문턱값(Threshold)은 사용자의 판단에 따라 다른 값으로 설정될 수 있다.

YOLO의 두 번째 버전부터는 여러 개의 앵커 박스(Anchor box)를 기반으로 이미지의 영역을 결정하게 되는데, 이때 IoU 값이 가장 높은 앵커 박스를 사용한다.

하지만 객체 검출에서 평균 제곱 오차(Mean squared error)나 평균 절대 오차(Mean absolute error)와 같은 손실 함수를 IoU 기반으로 대체하려는 시도는 많이 이루어지지 않는데, 예측값이 실제 값보다 클 경우와 이와 반대되는 경우를 가정해보면, 일반적이 손실 함수 값은 같지만 이를 IoU로 계산하면 값이 천차만별일 수 있다. 이처럼 손실 함수를 최소화하는 것과 IoU를 높이는 것 사이의 상관관계가 거의 존재하지 않기 때문에 이를 직접 사용하는 것은 크게 고려되지 않았다.

이러한 문제점을 보완하고 학습에 활용하기 위해 DIoU(Distance IoU)[8]와 같은 다양한 IoU 기반 손실 함수가 제안되고 있고, 이와 같은 시도가 점차 늘어나는 추세이다.

2.3. YOLO(You only look once)

대부분의 사람은 어떤 이미지를 봤을 때 이미지 내부에 있는 객체를 한눈에 파악할 수 있다. 하지만 비교적 최근에 제안된 Mask R-CNN[9]과 같은 계열의 검출 시스템은 복잡한 처리 과정으로 인해 이러한 '인간의 시각 체계'를 모방하기에는 부족한 부분이 많다.

YOLO는 앞서 언급한 인간의 시각 체계의 개념에 착안하여 이를 단일 회귀 문제(Single regression problem)로 간주하고, 단일 합성 곱 신경망을 통해 다중 경계 상자에 대한 확률을 계산하는 방식이다.

기존 객체 검출 방법보다 YOLO가 가지는 상대적인 장점은 처리 과정이 간단하여 속도가 매우 빠르고, 인식할 객체에 대해 좀 더 일반화된 특징을 학습한다는 것이다. 하지만 단순한 처리 과정으로 다른 검출 방법보다 상대적으로 정확도가 낮다는 단점이 있다. 또한, 입력 이미지를 고정된 하나의 크기가 아닌 다양한 크기로 변형하여 학습하기 때문에 여러 가지 해상도의 입력 데이터를 소화할 수 있다.

Ⅲ. 객체 검출 모델

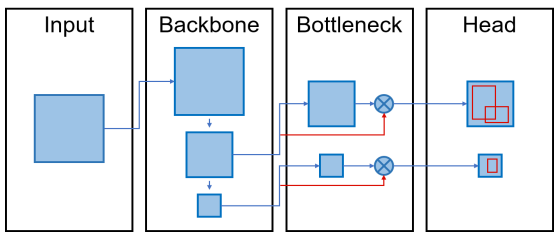

Fig. 2 Proposed model overview.

최근 제안된 검출 모델은 대부분 백본과 헤드 두 가지 부분이 연결된 구조를 취하고 있다. 백본은 VGGNet[10], ResNet[11], DenseNet[12]과 같이 입력 이미지의 특징을 추출하는 부분을 말하며, 헤드는 이렇게 추출된 특징을 기반으로 입력 내의 객체에 대한 클래스(Class)와 경계 상자(Bounding box)를 예측한다. 대표적으로 YOLO, RetinaNet[13] 등이 여기에 해당한다.

제안하는 모델은 이러한 백본, 헤드를 연결하는 병목층에 주의 집중 모듈을 끼워 넣어 성능 향상과 학습 시간 단축을 보이려 한다. 그림 2는 모델의 전체 구조를 도식화한 것으로, 백본에 Tiny Darknet, 병목 층에는 각종 주의 집중 기법(SENet, CBAM, ECA-Net), 그리고 헤드 부분에 YOLOv3를 적용하였다. 사용의 편의를 위해 각 부분은 전부 모듈화되어 원하는 부분을 손쉽게 다른 구조로 변경할 수 있도록 구성되었다.

3.1. 백본, Tiny Darknet

Tiny Darknet은 YOLO의 백본 모델인 Darknet을 경량화한 구조로, 레퍼런스 모델보다 성능은 떨어지지만 빠르게 학습하여 사용해볼 수 있다는 장점이 있다. 표 1은 백본에 사용된 모델의 구조를 나타낸 것으로, 특징 추출을 위한 9개의 3×3 합성 곱 층과 계산량 감소 및 비선형성 증가를 위한 1개의 1×1 합성 곱 층으로 이루어져 있다. 그리고 해당 구조에 업샘플링(Upsampling)을 추가하여 두 가지 스케일의 특징에 대해 학습을 하고, 이를 기반으로 좀 더 높은 정확도를 가질 수 있도록 모델을 수정하였다.

Table. 1 Proposed backbone model structure.

Type	Structure
Input	416×416×3 image
Conv, Pool	3×3×16, stride: 1, pad: 1 , pool: max
Conv, Pool	3×3×32, stride: 1, pad: 1 , pool: max
Conv, Pool	3×3×64, stride: 1, pad: 1 , pool: max
Conv, Pool	3×3×128, stride: 1, pad: 1 , pool: max
Conv, Pool	3×3×256, stride: 1, pad: 1 , pool: max
Conv, Pool	3×3×512, stride: 1, pad: 1 , pool: max
Conv	3×3×1024, stride: 1, pad: 1
Conv	1×1×256, stride: 1, pad: 0
Conv	3×3×512, stride: 1, pad: 1
Upsample	scale: 2, mode: nearest
Conv	3×3×256, stride: 1, pad: 1

3.2. 병목 층, 주의 집중 모듈

병목 층의 주의 집중 모듈에서는 백본에서 추출된 특징을 입력으로 받아 가중치를 계산하고, 헤드에서 결과를 예측할 때 가중치를 부여한다. 병목 층에서 다음 층으로 넘어가기 전, 즉, 정보량이 줄기 전에 모듈을 추가하여 중요한 부분의 값을 키우고, 덜 중요한 부분의 값을 줄이는 것이 핵심이며, 적은 연산량 증가로도 큰 성능 향상을 보이는 것이 목적이다.

본 연구에서 사용한 세 가지 기법 모두 단순한 합성곱과 풀링(Pooling)으로 이루어져 있으며, 어떤 합성 곱 신경망에도 쉽게 끼워 넣을 수 있도록 설계되었다.

3.3. 손실 함수, GIoU

기존의 YOLO는 자체적으로 고안한 손실 함수를 사용하고 있는데, 이는 매우 복잡하여 사용과 구현이 까다롭다. 이를 대체하기 위해 IoU를 직접 손실 함수로 사용할 수도 있지만, 앞선 장에서 언급했듯이 여기에는 많은 문제점이 뒤따른다. 또 다른 예로, 만약 실제값과 예측값 사이에 겹치는 부분이 없다면 IoU는 0이 될 것이고, 두 객체가 서로 얼마나 떨어져 있는지에 대한 부분이 학습에 전혀 반영되지 않을 것이다. 때문에 IoU를 손실함수로 사용하는 것에는 무리가 있다. GIoU[14]는 이러한 문제점을 보완하기 위해 제안된 방법으로 다음과 같이 계산된다.

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

$$GIoU = IoU - \frac{|C \setminus (A \cup B)|}{|C|} \tag{2}$$

$$L_{IoU} = 1 - IoU, \quad G_{GIoU} = 1 - GIoU \tag{3}$$

우선 비교할 두 영역 A, B에 대해 이를 모두 포함하는 가장 작은 영역인 C를 구한다. 그리고 식 1과 같이 A, B의 IoU를 계산한 다음, 식 2와 같이 GIoU를 계산한다. 식 2를 좀 더 자세히 살펴보면, C에서 A, B의 합집합을 뺀 후에 다시 C로 나누는데, 이는 A, B 사이의 빈 공간을 표준화하여 나타낸 것이다. 여기에 식 3과 같이 1에서 이 값을 빼고, 손실 함수로 사용한다.

그림 3은 IoU가 0으로 계산되는 각기 다른 상황에서, 이를 GIoU로 계산하면 어떠한 차이가 발생하는지를 보여준다. 여기서 볼 수 있듯이 두 영역의 빈 공간에 대한

정보까지 계산에 반영되면서, 손실함수로 사용하기에도 무리가 없다. 또한, 일반적인 IoU에 비해 수렴 속도가 빠르다는 것도 실험적으로 증명되었다.

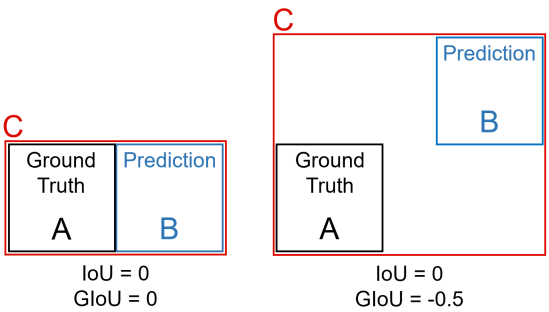

Fig. 3 Comparison of IoU and GIoU.

Ⅳ. 실험 및 결과

4.1. 실험 환경 및 설계

제안하는 모델은 윈도우(Microsoft windows) 운영체제 환경에서 파이썬(Python)으로 구현, 학습되었으며, 이를 위해 토치(Torch)기반의 오픈소스 머신 러닝 라이브러리인 파이토치(PyTorch)를 사용하였다. 또한, 일반적인 사용자 환경을 가정하여 단일 GPU만을 사용해 학습하였다.

학습 시 가중치 최적화 기법으로는 Nadam[15]을 사용하였는데, RMSprop에 모멘텀(Momentum)이 적용된 Adam[16]과는 달리, Nadam은 일반적인 모멘텀 대신 네스테로프 모멘텀(Nesterov momentum)을 적용하여 좀 더 빠르게 전역 최저점(Global minimum)을 찾을 수 있다는 장점이 있다.

학습 데이터는 9,163장의 학습 데이터 세트, 2,031장의 검증 데이터 세트, 그리고 821장의 테스트 데이터 세트로 이루어진 옥스퍼드 대학교의 Hand Dataset[17]을 사용하였으며, 총 800세대(Epoch)동안 학습을 진행하였다.

4.2. 평가 방법

대부분의 객체 검출 모델 성능은 정밀도-재현율 곡선(Precision-recall curve)과 평균 정밀도(Average precision)로 평가되는데, 여기서 정밀도는 모든 검출 결과 중 옳게 검출한 비율을 의미하며, 재현율은 명확히 검출 해내야 하는 객체 중 제대로 검출된 것의 비율을 의미한다. 이를 위해서는 객체가 제대로 검출되었을 때와 아닐 때를 구분할 수 있어야 하는데, 이때 기준으로 사용되는 것이 앞서 설명한 IoU이다.

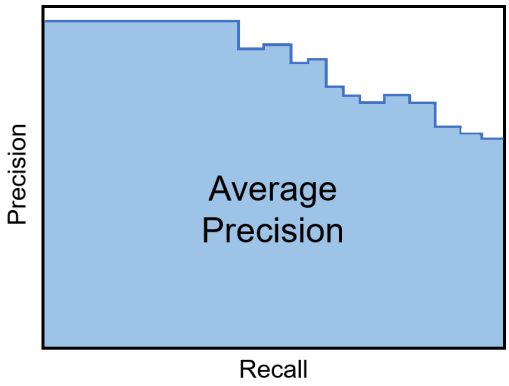

Fig. 4 Average precision.

정밀도-재현율 곡선이 전반적인 모델 성능을 파악하기에는 좋으나, 서로 다른 모델을 정량적으로 비교하기에는 불편한 점이 많다. 이를 해결하기 위해 평균 정밀도가 도입되었으며, 평균 정밀도는 문자 그대로 각 재현율 값(0에서 1까지)에 대응하는 정밀도의 평균이다. 즉, 그림 4와 같은 정밀도-재현율 그래프의 아랫부분에 해당한다.

본 연구에서는 성능 측면에서 평균 정밀도와 속도 측면에서 최적 성능으로의 수렴 속도를 평가지표로 사용하였다.

4.3. 실험 결과

표 2와 3은 위와 같이 설계된 실험 환경에서 YOLO를 단독 사용했을 때와 주의 집중 기법을 적용하였을 때, 결과에 어떠한 차이가 있는지를 보여준다.

Table. 2 Performance comparison.

Model	Params	AP [%]
Tiny YOLO	8.67M	60.1
Tiny YOLO + SENet	8.93M	61.8
Tiny YOLO + CBAM	8.81M	62.5
Tiny YOLO + ECA-Net	8.67M	62.2

Table. 3 Convergence time comparison.

Model	Convergence time
Tiny YOLO	778 Epochs, 15 Hours
Tiny YOLO + SENet	158 Epochs, 3 Hours
Tiny YOLO + CBAM	325 Epochs, 6 Hours
Tiny YOLO + ECA-Net	228 Epochs, 4 Hours

세 가지 주의 집중 기법 모두 단독 사용과 비교해 일정 부분 성능 향상을 보였으며, CBAM, ECA-Net, SENet 순으로 성능이 우수한 것으로 나타났다. 이때 오버헤드는 3% 미만으로, 학습 파라미터를 많이 증가시키지 않으면서도 검출 성능에 긍정적인 영향을 미침을 확인하였다.

표 3은 총 800세대 학습 중 최적 성능으로의 수렴 시간을 나타낸 것으로, 일반적인 Tiny YOLO 모델이 거의 마지막 세대에 이르러 최적 성능으로 수렴한 것과는 달리, 주의 집중 기법이 적용된 모델은 설정한 학습 세대의 절반도 지나지 않아 최적 성능으로 수렴한 것을 확인할 수 있다.

위 결과로 보아 주의 집중 기법을 적용한 모델 모두 성능과 수렴 속도에 이점이 있고, 특히 최적 성능으로의 수렴 속도가 단독 모델에 비해 월등히 빠른 것으로 나타났다.

V. 결 론

기계 학습에서 심층 학습으로, 신경망의 깊이가 깊어지고 모델의 구조가 복잡해질수록 성능은 비약적으로 향상되었지만, 요구되는 컴퓨팅 자원과 시간도 증가하게 되었다. 특히, 기계 학습을 활용한 이미지 처리 연산은 그 규모에 따라 정도의 차이는 있겠지만 최소 이틀에서 많게는 일주일 이상 학습하는 경우가 많았다. 컴퓨팅 자원을 늘리면 시간에 대한 문제는 어느 정도 해결되지만, 이 또한 임시방편일 뿐이다.

본 논문에서는 합성 곱 신경망에 주의 집중 및 다양한 최신 기법을 적용하여, 객체 검출에 활용할 수 있는 모델을 설계하였다. 제안하는 모델이 기존 모델보다 성능 향상은 물론 학습 시간을 현저히 낮출 수 있음을 실험적으로 증명하였으며, 이를 통해 일반적인 사용 환경에서의 기계 학습에 대한 접근성 증가가 기대된다.

ACKNOWLEDGEMENT

This work was supported by the Dongseo University Research Fund of 2020. (DSU-20200017)

REFERENCES

[1] J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement,” University of Washington, Washington: WA, Technical Report, 2018.

[2] K. Xu, J. Ba, R. Kiros, K. Cho, and A. Courville, “Show, attend and tell: Neural image caption generation with visual attention,” in *International conference on machine learning*, France: FR, pp. 2048-2057, 2015.

[3] H. Nam, J. Ha, and J. Kim, “Dual attention networks for multimodal reasoning and matching,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii: HI, pp. 299-307, 2017.

[4] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, and X. Tang, “Residual attention network for image classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii: HI, pp. 3156-3164, 2017.

[5] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Utah: UT, pp. 7132-7141, 2018.

[6] S. Woo, J. Park, J. Lee, and K. So, “Convolutional block attention module,” in *Proceedings of the European conference on computer vision (ECCV)*, Germany: DE, pp. 3-19, 2018.

[7] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, “ECA-net: Efficient channel attention for deep convolutional neural networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Virtual, pp. 11534-11542, 2020.

[8] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, “Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression,” in *Proceeding of the AAAI Conference on Artificial Intelligence*, New York: NY, vol. 34, no. 7, pp. 12993-13000, 2020.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii: HI, pp. 2961-2969, 2017.

[10] H. Qassim, A. Verma, and D. Feinzimer, “Compressed residual-VGG16 CNN model for big data places image recognition,” in *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, Nevada: NV, pp. 169-175, 2018.

[11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Nevada: NV, pp. 770-778, 2016.

[12] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii: HI, pp. 4700-4708, 2017.

[13] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii: HI, pp. 2980-2988, 2017.

[14] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, California: CA, pp. 658-666, 2019.

[15] T. Dozat, “Incorporating nesterov momentum into adam,” in *ICLR 2016 workshop submission*, Puerto Rico: PR, 2016.

[16] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, California: CA, pp. 1-15, 2015.

[17] A. Mittal, A. Zisserman, and P. Torr. Hand Dataset [Internet]. Available: http://www.robots.ox.ac.uk/~vgg/data/hands/.

김근식(Geun-Sik Kim)

단국대학교 일본어과 학사
단국대학교 경영학과 학사
부산대학교 정보컴퓨터공학부 학사
부산대학교 정보융합공학과 석사과정 (현재)
※관심분야 : 컴퓨터 비전, 이미지 캡셔닝, 회귀 분석, 심층 학습

배정수(Jung-Soo Bae)

부산대학교 해양과학과 학사
부산대학교 멀티미디어협동과정 석사
부산대학교 멀티미디어협동과정 박사
동서대학교 소프트웨어융합대학 조교수 (현재)
※관심분야 : 인공지능, 인공적 도덕 행위자

차의영(Eui-Young Cha)

서울대학교 자연대학 전자계산학 이학석사
서울대학교 공과대학 컴퓨터공학 박사
한국전자기술연구소 연구원
부산대학교 정보컴퓨터공학부 교수 (현재)
※관심분야 : 지능형 로봇, 자율 주행, 심층 학습