JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 24, No. 2: 204~211, Feb. 2020

주 객체 위치 검출을 위한 Grad-CAM 기반의 딥러닝 네트워크

김선진[1] · 이종근[1] · 곽내정[2] · 류성필[2] · 안재형[3*]

Grad-CAM based deep learning network for location detection of the main object

Seon-Jin Kim[1] · Jong-Keun Lee[1] · Nae-Jung Kwak[2] · Sung-Pil Ryu[2] · Jae-Hyeong Ahn[3*]

[1]Graduate Student, Department of Information and Communication Engineering, Chung-buk National University, Cheong-ju, 28664 Korea
[2]Lecturer, Department of Information and Communication Engineering, Chung-buk National University, Cheong-ju, 28664 Korea
[3*]Professor, Department of Information and Communication Engineering, Chung-buk National University, Cheong-ju, 28664 Korea

요 약

본 논문에서는 약한 지도학습을 통한 주 객체 위치 검출을 위한 최적의 딥러닝 네트워크 구조를 제안한다. 제안된 네트워크는 약한 지도학습을 통한 주 객체의 위치 검출 정확도를 향상시키기 위해 컨벌루션 블록을 추가하였다. 추가적인 딥러닝 네트워크는 VGG-16을 기반으로 합성곱 층을 더해주는 5가지 추가적인 블록으로 구성되며 객체의 실제 위치 정보가 필요하지 않는 약한 지도 학습의 방법으로 학습하였다. 또한 객체의 위치 검출에는 약한 지도학습의 방법 중, CAM에서 GAP이 필요하다는 단점을 보완한 Grad-CAM을 사용하였다. 제안한 네트워크는 CUB-200-2011 데이터 셋을 이용하여 성능을 테스트하였으며 Top-1 Localization Error를 산출하였을 때 50.13%의 결과를 얻을 수 있었다. 또한 제안한 네트워크는 기존의 방법보다 주 객체를 검출하는데 더 높은 정확도를 보인다.

ABSTRACT

In this paper, we propose an optimal deep learning network architecture for main object location detection through weak supervised learning. The proposed network adds convolution blocks for improving the localization accuracy of the main object through weakly-supervised learning. The additional deep learning network consists of five additional blocks that add a composite product layer based on VGG-16. And the proposed network was trained by the method of weakly-supervised learning that does not require real location information for objects. In addition, Grad-CAM to compensate for the weakness of GAP in CAM, which is one of weak supervised learning methods, was used. The proposed network was tested through the CUB-200-2011 data set, we could obtain 50.13% in top-1 localization error. Also, the proposed network shows higher accuracy in detecting the main object than the existing method.

키워드 : 객체 검출, 딥러닝, 약한 지도학습, VGG-16

Keywords : Deep learning, Object detection, VGG-16, Weakly-supervised learning

Received 13 December 2019, **Revised** 21 December 2019, **Accepted** 24 December 2019

*** Corresponding Author** Jae-Hyeong Ahn(E-mail:jhahn@chungbuk.ac.kr, Tel:+82-43-261-2483)
Professor, Department of Information and Communication Engineering, Chungbuk National University, Cheongju, 28664 Korea

Open Access http://doi.org/10.6109/jkiice.2020.24.2.204 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

최근 딥러닝이 발전하면서 딥러닝 모델을 활용한 객체 인식이나 물체의 위치 검출 등의 분야에서 딥러닝을 이용한 연구가 활발하다. 실제로 ILSVRC(Image Large Scale Visual Recognition Challenge)[1]와 같은 대회에서도 다양한 딥러닝을 활용한 객체 인식 방법들이 상위권을 차지하고 있다.

딥러닝을 통한 객체 인식 방법 중, 특히 faster-R--CNN[2]이나 SSD[3]와 같은 방법은 그중에서도 뛰어난 성능을 보여준다. 위의 딥러닝 방법들은 이미 만들어진 데이터 셋과 그 안에 포함되어있는 객체의 위치에 대한 실제 정보를 같이 학습하는 방법으로 설계되어 있는데, 이런 딥러닝 학습 모델들을 완전 지도학습 방법이라고 한다.

완전 지도학습 방법을 통한 객체의 위치 검출 방법은 성능이 뛰어나지만, 물체의 위치에 대한 실제 정보를 학습 과정에 반드시 포함해야 한다는 단점이 있는데, 이 때문에 시간이 지날수록 다양한 데이터를 학습하면서 객체의 위치에 대한 레이블을 만들어 주어야 하는 데 많은 시간을 소모해야 하는 문제가 있다.

그래서 최근 연구자들은 완전 지도학습 방법 외에 약한 지도학습(Weakly-supervised learning)의 방법을 통해 다양한 연구를 하고 있다. 약한 지도학습이란 학습 과정에서 이미지와 그에 대한 클래스 레이블만을 학습시키는 방법이다. 물론 많은 연구에서 분류나 위치 검출에 대한 효율은 비교적으로 많은 정보를 학습한 완전 지도학습의 경우가 더 좋다고 알려져 있다. 하지만 완전 지도학습과 비교하여 약한 지도학습은 객체의 실제 위치에 대한 레이블이 필요하지 않기 때문에 많은 인적, 물적 낭비를 줄일 수 있다는 장점이 있다. [4, 5]

약한 지도학습의 객체 인식 방법 중 CAM(Class Activation mapping)[6]과 Grad-CAM(Gradient-based Class Activation Mapping)[7]은 객체들의 차별적인 특징들을 컨벌루션 층에서 추출하고 시각화하여 객체를 찾아주는 방법으로써 최근 다양한 연구에서 활용되고 있다. 그 중 Choe와 Shim[4], Singh과 Lee[8]의 연구에서는 CAM의 방법을 활용하여 약한 지도학습의 객체 인식 정확도를 높이는 방법을 제안하고 있다. 하지만 CAM의 방법을 활용하는 방법은 기존의 CNN(Convol--ution Neural Network) 구조에서 활용하던 완전 연결 계층(Fully-Connected layer)을 제거하고 NIN[9]에서 제안한 GAP(Global Average Pooling)을 삽입하여 기존의 구조를 변경해야 한다는 단점이 있다.

Grad-CAM은 이러한 CAM의 구조를 변경해야하는 단점을 보완한 약한 지도학습의 객체 인식 방법이다. 하지만 Grad-CAM과 CAM 모두 딥러닝 모델의 네트워크 구조에 따라 객체 인식의 효율에 차이를 보인다. 또한 두 방법 모두 입력 이미지에서 가장 차별적인 특징에 초점을 맞추기 때문에 객체 인식 정확도에서 그 정확도가 높지 않을 수 있다.

따라서 본 논문에서는 CAM의 방법 대신 Grad-CAM을 활용하여 주 객체의 인식을 통하여 약한 지도학습을 통한 객체 인식의 정확도를 향상하고자 한다. 본 논문에서는 VGG-16[10] 구조에서 추가적인 컨벌루션 블록을 삽입하여 딥러닝 네트워크 구조에서 컨벌루션 층을 늘리고 학습 과정에서 더욱 많은 특징을 보게 하여 약한 지도학습을 통한 객체 인식의 정확도를 향상하고자 한다. 이러한 방법을 통한 주 객체 인식의 정확도 향상 방법은 다른 연구에서보다 간단히 구현할 수 있으며, 본 논문에서 제안하는 방법은 다른 네트워크와 간단히 결합하여 객체 검출 정확도에 향상을 가져올 수 있다.

Ⅱ. 관련 연구

2.1. Grad-CAM[7]

Grad-CAM[7]은 CAM[6]의 일반화된 방법으로 CAM에서의 단점을 보완한 방법이다. CAM에서의 가장 큰 제약점은 GAP(Global Average Pooling)[9]의 구조가 필수적이기 때문에 네트워크 구조를 변경해야 한다는 것이다. 하지만 Grad-CAM은 일련의 과정 없이 기존의 모든 CNN 구조에서도 사용할 수 있다.

$$a_k^c = \frac{1}{z}\sum_i\sum_j \frac{\delta y^c}{\delta A_{ij}^k} \tag{1}$$

$$L_{Grad-CAM}^c = ReLU(\sum_k a_k^c A^k) \tag{2}$$

수식 1,2는 Grad-CAM을 나타낸다. 수식 1는 각 특징 맵에서의 특징에 대한 중요도를 나타낸다. 이 공식은 각 특징 맵에서 입력 값에 대한 미분 값을 나타낸다. 각 특징 맵 픽셀에서의 중요도 값을 통해 Grad-CAM을 계산

할 수 있는데 이는 수식 2를 통해 볼 수 있다.

Grad-CAM의 또 다른 장점은 이 시각화에 대한 값을 어떠한 컨벌루션 층에서도 계산할 수 있다는 것이다. CAM은 GAP을 사용하였기 때문에 시각화 값은 마지막 컨벌루션 층에만 적용할 수 있는 데 비하여, Grad-CAM은 각 컨벌루션 층의 특징 맵에서 미분을 통하여 중요도를 계산하기 때문에 어떠한 컨벌루션 층에서도 사용할 수 있다.

하지만 CAM과 Grad-CAM은 시각화 값을 추출하는 딥러닝 모델의 네트워크 구조에 따라 객체 인식의 효율에 차이를 보인다. 또한, 일반적인 합성곱 신경망에서 분류기가 분류 정확도를 높이기 위하여 가장 차별적인 특징에 초점을 맞추기 때문에 기존의 CAM과 Grad-CAM을 통해 객체 인식 효율을 구하게 되면 그 정확도가 높지 않을 수 있다[4].

2.2. Weakly Supervised Object Localization

약한 지도학습이란, 학습 과정에서 이미지와 그에 대한 클래스 레이블만을 학습시켜 딥러닝 예측 모델을 생성하는 방법이다. 약한 지도학습에서의 객체 검출 방법은 최근 Choe와 Shim[4], Singh과 Lee[8] 등의 다양한 연구에서 찾아볼 수 있다. Choe와 Shim의 연구에서는 2.2절에서 언급한 CAM과 Grad-CAM의 합성곱 신경망에서의 분류기가 가장 차별적인 특징에 초점을 맞추기 때문에 객체 검출 효율이 떨어지는 것에 대한 문제를 제기하고 있다.

그 중 Choe와 Shim의 연구에서는 Singh과 Lee의 연구에서 보여준 드롭-마스크(Drop-mask)에 중요도-맵(Importance map)을 결합하여 ADL(Attention-based Dropout Layer)을 제시하고 있다. 하지만 Choe와 Shim의 연구에서도 Singh과 Lee의 연구에서와 마찬가지로 CAM의 방식을 활용하여 기존의 CNN 구조에서 완전 연결 계층을 제거하고 GAP을 삽입하여 네트워크 구조를 재구성하여야 한다는 단점이 존재한다.

Ⅲ. 주 객체 검출을 위한 Grad-CAM 기반의 딥러닝 네트워크

객체 인식의 경우 많은 특징이 분류기에 전달될수록 객체 위치 검출의 정확도가 높다. 그러므로 Choe와 Shim, Singh과 Lee의 연구에서는 기존의 완전 연결 계층 대신 더 많은 특징을 전달할 수 있는 GAP을 활용하고 이를 활용한 CAM의 방법을 통해 약한 지도학습에서의 주 객체 위치 검출 정확도를 높이는 방법을 제안하고 있다. 하지만 GAP을 활용하기 위해서는 기존의 완전 연결 계층을 제거하고 GAP을 삽입하는 일련의 과정을 추가로 수행해야 하는 단점이 있다.

따라서 본 논문에서는 약한 지도학습을 통한 주 객체의 위치 검출 방법 중, 기존의 CNN 구조를 변경하지 않는 Grad-CAM의 방법을 활용하여 주 객체의 위치를 검출한다. 또한 VGGNet[10]의 연구 결과를 기반으로 객체 위치 검출 정확도를 높이기 위해 VGGNet에 컨벌루션 층을 추가하여 그 성능을 분석하고 약한 지도학습을 통한 주 객체 위치 검출을 위한 최적의 네트워크 구조를 구성한다.

Table. 1 Configuration of proposed additional blocks

Block A	Block B	Block C	Block D	Block E
Conv 3x3	Conv 3x3	Conv 3x3	Conv 3x3	Conv 3x3
	Conv 3x3	Conv 3x3	Conv 3x3	Conv 3x3
		Conv 3x3	Conv 3x3	Conv 3x3
			Conv 3x3	Conv 3x3
				Conv 3x3
Max-Pooling	Max-Pooling	Max-Pooling	Max-Pooling	Max-Pooling

3.1. 객체 검출 정확도를 높이기 위한 블록이 추가된 네트워크 모델

본 논문에서는 서로 다른 깊이의 컨벌루션 층을 가진 5개의 블록을 설계하여 컨벌루션 층의 깊이에 따른 주 객체 위치 검출 효율을 비교하며 5개 블록의 결과의 비교로 주 객체 검출 효율이 가장 높은 블록을 찾는다. 서로 다른 5개의 블록은 1개부터 5개의 컨벌루션 층을 추가로 가지며, 컨벌루션 층의 커널은 3x3의 크기로 고정한다. 이때 공통으로 1개의 최대 풀링 층을 가지게 설계하였는데, 5개의 블록이 공통으로 1개의 최대 풀링 층을 갖는 이유는 컨벌루션 층의 깊이 외에 다른 변수를 차단하여 컨벌루션 층의 깊이와 객체 검출 효율의 상관관계

를 분석하기 위한 것이다. 표 1은 본 논문에서 제안하는 추가적인 블록의 구성을 보여준다.

블록을 추가로 사용하는 이유는 학습 과정에서 더 많은 특징을 보여주어 주 객체의 위치 검출 효율을 늘려주는 목적 외에도 기존의 CNN 구조를 통한 분류 정확도를 유지해주기 때문이다. VGGNet[10]의 연구 결과를 보면 컨벌루션 층의 깊이를 단순히 증가시키는 방법에 따라, 분류 정확도를 유지 혹은 상승시키는 효과가 있다고 하였다.

이에 따라, 제안하는 추가적인 블록은 컨벌루션 층의 깊이를 깊게 하여 기존의 CNN 구조를 통한 분류 정확도에 영향을 미치지 않게 한다. 또한, 추가된 블록을 사용하면 학습 과정에서 입력 이미지의 다양한 특징을 학습하기 때문에, 주 객체 검출 정확도는 증가시키는 효과가 있다.

결론적으로 제안하는 추가적인 블록은 기존의 사전 학습된 VGG-16에서 완전 연결 계층 이전에 추가로 삽입되고, 컨벌루션 층의 깊이를 늘려 기존의 구조보다 더욱 많은 특징을 학습한 결과를 출력한다.

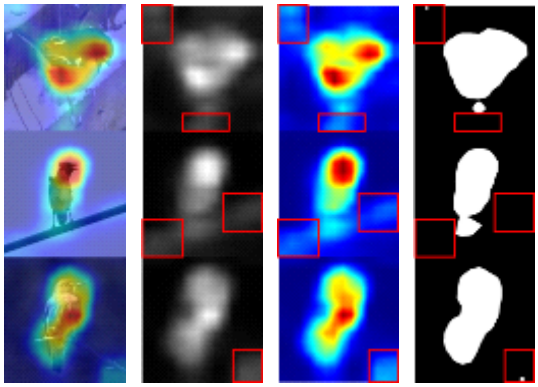
3.2. 객체 검출 방법 및 B-Box 데이터 생성

제안하는 블록을 추가한 딥러닝 네트워크는 기본적인 합성곱 신경망의 구조를 가진다. 이런 합성곱 신경망의 구조에서 주 객체의 위치를 검출하기 위하여 본 논문에서는 CAM의 일반화된 방법인 Grad-CAM을 사용한다.

Grad-CAM을 검출한 후 출력된 이미지는 최적화된 임계값 수치를 통해 임계값 처리 과정을 거친다. 이전 연구에서 Zhang 등[11]은 최적화된 임계값 처리는 객체 위치 검출 효율의 향상을 가져올 수 있다고 하였다. 너무 큰 임계값을 가지면 유용한 영역 발견을 효과적으로 유도할 수 없고 너무 작은 임계값을 가지면 영상에서 잡음이 섞일 수 있다고 하였다. 그림 1은 CUB-200-20- -11에서 Grad-CAM을 주고 임계값 처리를 하였을 때의 영상을 보여 준다.

그림 1의 (b), (c)에서 표시한 영역을 보면 필요한 영역 외의 부분까지 Grad-CAM을 통해 검출되는 것을 알 수 있다. 이러한 영역을 제거하기 위해 임계값 처리한 부분은 그림 1의 (d)에서 볼 수 있는데 그림 1의 (b), (c)에서의 불필요한 부분을 효과적으로 제거하고 필요한 부분은 효과적으로 검출하는 것을 볼 수 있다.

따라서 본 논문에서는 먼저 입력 영상에서 객체를 Grad-CAM을 통해 검출하고, 최적화된 임계값 수치를 찾아내어 B-Box(Bounding Box)를 생성하여 성능 평가에 필요한 데이터를 생성하였다.



(a) (b) (c) (d)

Fig. 1 The image that removes unwanted parts through the grad-CAM image and threshold processing.

Ⅳ. 실험 및 결과 분석

4.1. 실험 환경

본 논문의 실험은 윈도우 환경에서 진행하였다. PC 메모리는 12GB이며, 딥러닝을 위한 GPU를 따로 사용하였다. GPU를 위한 그래픽 카드는 NVIDIA Geforce GTX1050Ti를 사용했으며 그래픽 메모리는 4GB이다.

딥러닝 모델 학습과 예측 모두 Python으로 진행하였다. Python에서 딥러닝을 위해 Tensorflow와 Keras 라이브러리를 병행하여 사용하였다.

4.2. 데이터 셋

본 논문에서는 CUB-200-2011(Caltech-USCD Bird-s-200-2011)[12]을 데이터를 활용하였다. CUB-200-2--011은 각 종의 새들로 이루어진 200가지 클래스 레이블로 이루어져 있으며 물체의 위치에 대한 실제 정보를 포함하고 있다. 학습 과정에서는 두 가지 데이터 셋 모두 이미지 데이터와 클래스 레이블만을 활용하여 약한 지도학습 방법으로 학습하였다. 물체의 위치에 대한 실제 정보 레이블은 평가 과정에서만 사용한다.

4.3. 성능 평가 방법

본 논문에서는 성능 평가 방법을 위해 ILSVRC에서 제안한 이미지 분류 효율 알고리즘과 객체 위치 검출 효율 알고리즘을 사용한다[1]. 해당 성능 평가 방법은 다양한 연구에서 활용되고 있으며, ILSVRC에서 사용된 객체의 위치 검출 에러 측정 방식은 기존의 IoU 방식보다 명확한 기준을 가진 방식이다. 수식 3은 이미지 데이터 셋에 따른 분류 성능을 산출하는 알고리즘으로 이미지 분류에 대한 성능은 합성곱 신경망을 통해 생성된 예측 모델을 통하여 분류된 이미지들의 개수를 통하여 다음과 같이 산출한다.

$$c.r = \frac{\text{옳게 분류된 이미지의 수}}{\text{전체 이미지의 수}} \tag{3}$$

여기서 $c.r$은 전체 이미지의 수에 대한 옳게 분류된 이미지의 수의 비율을 나타낸다. 수식 4,5는 객체의 위치 검출 효율의 에러를 측정하는 수식이다.

$$e = \min_i(\min_j(\max(d_{ij}, f_{ij}))) \tag{4}$$

$$e.r = \frac{\sum e}{\text{전체이미지의수}} \tag{5}$$

수식 4에서 d는 클래스에 대한 레이블이 예측된 결과와 같으면 0을, 아니면 1의 값을 가진다. f는 예측된 B-Box와 객체의 위치에 대한 실제 정보 B-Box가 50% 이상 겹치면 0을, 아니면 1의 값을 가진다. i 와j는 각각 예측한 B-Box와 실제 정보에 포함된 B-Box의 인덱스를 나타내며, e는 그를 통해 산출된 하나의 이미지에 대한 에러 카운트이다. 즉 데이터 셋에서 객체의 위치 검출 에러율은 수식 5와 같다.

4.4. 최적화된 임계값 검출

본 논문에서는 주 객체의 위치 검출에 있어서 Grad-CAM을 검출한 이후 임계값 처리를 수행한다. Zhang 등[11]은 너무 큰 값으로 임계값 처리를 하게 되면 유용한 부분을 검출하기 힘들고, 너무 작은 값으로 임계값 처리를 하게 되면 영상에 잡음이 섞일 수 있다고 하였다. 따라서 본 논문에서는 주 객체의 위치 검출의 효율을 최대화하기 위하여 최적의 임계 값을 찾았다.

표 2는 임계값에 대한 실험 결과를 나타내는 표 이다. 최적화된 임계값을 검출하기 위한 실험은 CUB-20--0-2011 데이터 셋을 사용하였으며, Grad-CAM을 통해 검출한 영상에서의 최대 픽셀에 따라 임계값을 설정하여 진행하였다. 먼저 임계값을 10%~35%의 범위에서 5%의 간격을 두고 실험을 진행한 경우, 20%의 임계값을 주었을 때 35.623%의 주 객체 위치 검출을 보이며 가장 높은 효율을 보였다. 이후 20% 임계값에서 ∓2% 범위를 추가로 설정하고 1% 간격을 두고 추가로 실험을 진행하였다. 실험 결과를 통해 19% 임계값을 주었을 때 가장 높은 효율을 보임을 알 수 있었다.

Table. 2 Top-1 Localization Error of the main object by threshold processing

Threshold(%)	Top-1 Loc. error(%)
10	67.981
15	64.546
19	**64.249**
20	64.377
21	64.631
25	67.939
30	72.731
35	76.336

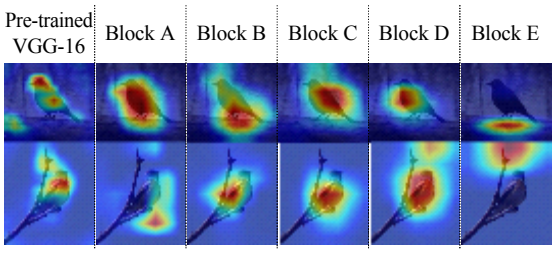


Fig. 2 Grad-CAM extracted from CUB-200-2011 data sets according to the experimental process

4.5. 이미지 분류 및 객체 검출

본 논문에서는 이미지 분류와 Grad-CAM을 적용한 객체 검출 성능을 확인하기 위하여 사전 훈련된 VGG-16과 표 1의 추가적인 블록을 삽입한 5개의 네트워크의 총 6가지의 네트워크 모델을 구성하고 이미지 분류 정확도와 객체 위치 검출 정확도를 평가한다.

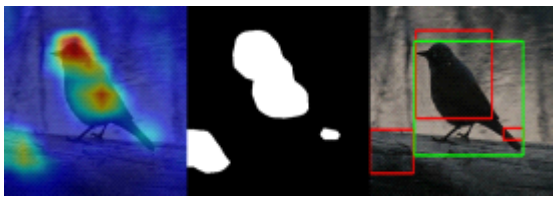
그림 2는 제안 방법에 따라 사전 학습된 VGG-16에 표 1의 추가적인 블록을 삽입하여 학습한 뒤 Grad-CAM을 적용한 결과이다. 여기서 Pre-trained VGG-16은 기존의 사전 학습된 VGG-16을 학습하여 나온 결과이며, Block A부터 Block E까지는 각각 표 1의 추가적인 블록

을 삽입하여 추출한 결과이다.

그림 2의 (a), (b)에서 Pre-trained VGG-16을 보면 부리나 머리 부분에 Grad-CAM이 모여 있는 것을 볼 수 있다. 이것은 Grad-CAM이 입력된 이미지에서 가장 차별적인 특징에 주목한다는 것을 의미한다. 이후에 추가적인 블록을 사용하면 맨 처음 초점을 두었던 머리 부분을 포함한 상태에서 컨벌루션 층이 몸이나 다리 등과 같은 다른 특징을 찾아가는 것을 볼 수 있다.

그림 3은 Grad-CAM 결과를 통하여 임계값 처리를 하고 B-Box 데이터를 생성한 결과를 보여준다. 여기서 초록색 B-Box는 실제 정보에 존재하는 객체의 위치 정보에 기반을 두었으며, 빨간색 B-Box의 경우 예측된 객체의 위치 정보에 기반을 둔다.

그림 3의 (b), (d)에서 Block C 구조를 사용하였을 때 Block C에서 우리가 찾고자 하는 객체를 중심으로 B-Box가 생성된 것을 볼 수 있다. 하지만 그림 3의 (a), (c)에서 Pre-trained VGG-16의 경우 그 B-Box의 위치가 머리 부분으로, 가장 차별적인 부분을 중심으로 생성되어 있음을 알 수 있다.



(a) B-box data generation through Rusty-Blackbird in pre-trained VGG-16

(b) B-box data generation through Rusty-Blackbird in Block C

(c) B-box data generation through Lazuli-Bunting in pre-trained VGG-16

(d) B-box data generation through Lazuli-Bunting in Block C

Fig. 3 Results of B-Box data generation on Pre-trained VGG-16 and Block C on CUB-200-2011

표 3은 예측한 객체의 위치 정보에 따라 실제 정보와 비교한 실험의 평가표이다.

Table. 3 Evaluation of top-1 localization error and top-1 classification accuracy.

CUB-200 -2011	Top-1 Classification accuracy (%)	Top-1 Localization error (%)
VGG-16	62.42	64.25
VGG-16 + Block A	59.79	61.71
VGG-16 + Block B	60.56	52.42
VGG-16 + Block C	**60.51**	**50.13**
VGG-16 + Block D	61.74	52.80
VGG-16 + Block E	60.51	56.49

여기서 Top-1 Localization error의 경우 에러율로써 값이 낮을수록 그 성능이 더 좋음을 의미한다. Top-1 Classification accuracy는 정확도로써 값이 높을수록 그 성능이 좋다.

표 3의 실험 평가표에서 보면 CUB-200-2011의 경우 Block C를 사용하였을 때 그 주 객체의 위치 검출 효율이 가장 높다. 이후 가장 높은 효율이 나온 이후 주 객체의 위치 검출 효율이 점점 낮아지는 것을 알 수 있다. 이는 컨벌루션 층이 늘어나면 주 객체의 위치 검출 효율은 향상할 수 있지만, 그 한계가 있음을 알 수 있다.

분류 정확도를 보면 약간의 감소는 있지만, 기존의 VGG-16 구조를 활용했을 때와 큰 차이를 보이지 않는다. 이는 추가적인 블록을 사용함에 따라, 기존의 CNN 구조의 분류 정확도는 큰 영향을 받지 않는다는 것을 보여준다.

표 4는 본 논문에서 제안한 추가적인 블록 중 가장 높은 효율을 보여준 블록 C인 제안 방법과 Choe와 Shim[4]이 보여준 연구 결과를 비교한 표이다.

Table. 4 Comparison of results from Choe and Shim[4] with proposed method

CUB-200-2011	Top-1 Localization err.(%)
VGG-16	64.25
Proposed Method	**50.13**
VGG-GAP + ADL-A	65.59
VGG-GAP + ADL-B[4]	50.31

표 4에서 ADL-A와 B는 Choe와 Shim이 제안한 드롭-마스크와 중요도-맵의 비율에 따라 다른 결과를 보여준다. ADL-A는 드롭-마스크와 중요도-맵을 전부 사용하지 않았을 때의 결과이며 ADL-B는 Choe와 Shim의 연구에서 가장 좋은 주 객체 검출 효율을 보여준 드롭-마스크 75%, 중요도-맵 25%를 적용했을 때의 결과이다.

주 객체의 검출 효율을 보면 완전 연결 계층을 제거한 CAM을 사용한 Choe와 Shim의 방법보다 본 논문에서 제시한 기존의 VGG-16 구조에서 Block C를 추가한 Proposed Method를 사용했을 때 더 높은 정확도를 가지는 것을 알 수 있다.

Ⅴ. 결 론

본 논문에서는 약한 지도학습을 통한 주 객체의 위치 검출 효율을 높이는 방법으로 기존의 VGG-16 구조에 컨벌루션 층으로 이루어진 추가적인 블록을 제안하고 최적화된 딥러닝 모델(VGG16 + Block C)을 제시하였다. 이렇게 Grad-CAM에 최적화된 딥러닝 모델은 분류 정확도에 큰 영향을 미치지 않으며, 주 객체의 위치 검출 효율은 향상되는 결과를 보였다. 또한 제안된 네트워크를 구성하는 실험 방법을 통하여, 컨벌루션 층의 깊이에 따른 주 객체의 위치 검출 효율 향상과 그 한계점을 보여주었다. 또한, 기존의 CNN 구조에서 단순히 컨벌루션 층을 늘려주는 추가적인 블록을 사용함으로써 다양한 CNN 구조의 딥러닝 네트워크에서 사용할 수 있음을 보여주었다.

본 논문은 제안하는 추가적인 블록을 시간적, 물리적 제약으로 인하여 VGG-16의 네트워크 구조에서만 실험했으므로 향후 제안하는 방법을 ResNet[13]이나 GoogLeNet 등과 같은 다양한 CNN의 구조에서 적용하는 연구가 필요하다. 또한 Fickle-Net[14]의 연구에서 보여준 Drop-Out에 따른 주 객체의 위치 검출 정확도의 증가 방법에 기반을 두어 더 나은 정확도를 가져올 수 있을 것으로 기대된다.

References

[1] 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,” *arXiv:1409.0575v3*, 2015.

[2] S. Ren, K. He, R. Girshick, and J. Sun. “Faster R-CNN: towards real-time object detection with region proposal networks,” *arXiv:1506.01497v3*, 2016.

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, and A. C. Berg., “SSD: Single Shot MultiBox Detector,” *arXiv:1512.02325v5*, 2016.

[4] J. Choe, and H. Shim, “ADL:Attention-based Dropout Layer for Weakly Supervised Object Localization,” *arXiv:1908.10028v1*, 2019.

[5] Y. Wei, J. Feng, X. Liang, M. M. Cheng, Y. Zhao, and S. Yan, “Object region mining with adversarial erasing: A simple classification to semantic segmentation approach,” *arXiv:1703.08448v3*, 2018.

[6] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. “Learning Deep Features for Discriminative Localization,” *arXiv:1512.04150*, 2015.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” *arXiv: 1610.02391*, 2016.

[8] K. K. Singh, and Y. J. Lee, “Hide-and-Seek: Forcing a network to be meticulous for weakly-supervised object and action localization,” *arXiv:1704.04232v2*, 2017.

[9] M. Lin, Q. Chen, and S. Yan, “Network In Network,” *arXiv:1312.4400*, 2013.

[10] K. Simonyan, and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv: 1409.1556*, 2014.

[11] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. Huang, “Adversarial complementary learning for weakly supervised

object localization," *arXiv:1804.06962v1*, 2018.
[12] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," *California Institute of Technology*, 2011.
[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385*, 2015.
[14] J. Lee, E. Kim, S. Lee, J. Lee, and S. Yoon, "FickleNet: Weakly and Semi-supervised Semantic Image Segmentation using Stochastic Inference," *arXiv:1902.10421*, 2019.

김선진(Seon-Jin Kim)

2018년 2월 충북대학교 정보통신공학과 공학사
2018년 3월 ~ 현재 충북대학교 정보통신공학과 석사과정
※관심분야 : 영상처리, 딥러닝, 객체검출

이종근(Jong-Keun Lee)

2018년 2월 충북대학교 정보통신공학과 공학사
2018년 3월 ~ 현재 충북대학교 정보통신공학과 석사과정
※관심분야 : 영상처리, 멀티미디어 프로그래밍, 모바일 프록래밍, 머신러닝, 딥러닝

곽내정(Nae-Joung Kwak)

1995년 2월 충북대학교 정보통신공학과 공학석사
2005년 2월 충북대학교 정보통신공학과 공학박사
2005년 3월 ~ 2009년 2월 목원대학교 정보통신공학과 초빙교수
1995년 3월 ~ 현재 충북대학교 시간강사
※관심분야 : 영상처리, 멀티미디어 프로그래밍, 컴퓨터 비전, 머신러닝, 딥러닝

류성필(Sung-Pil Ryu)

2014년 2월 : 충북대학교 정보통신공학과(공학박사)
2004년 2월 ~ 2006년 3월 : LG 전자 단말연구소 주임연구원
2009년 1월 ~ 2013년 1월 : HERO Tech 기술이사
2006년 9월 ~ 현재 : 충북대학교 정보통신공학과 강사
※관심분야 : 멀티미디어 정보처리, MPEG, H.264, 컴퓨터 비전, 딥러닝.

안재형(Jae-Hyeong Ahn)

1981년 2월 : 충북대학교 전기공학과(공학사)
1983년 2월 : 한국과학기술원 전기 및 전자공학과(공학석사)
1991년 8월 : 한국과학기술원 전기 및 전자공학과(공학박사)
1987년 ~ 현재 : 충북대학교 정보통신공학부 교수
※관심분야 : 영상처리 및 영상 정보 처리, 멀티미디어 제작 및 정보 제공, 인터넷 통신 및 프로그래밍