Object Detection 성능 향상을 위한 피라미드 Squeeze-and-Excitation Block 활용 방법

장현아*, 최혜연*, 김범준*, 김상우*
포항공과대학교*

Pyramid Squeeze-and-Excitation Block with Transposed Convolution for Performance Improvement of Object Detection

Hyeonah Jang*, Hyeyeon Choi*, Bum Jun Kim*, Sang Woo Kim*
Pohang University of Science and Technology*

Abstract - 본 논문에서는 PFPNet을 기반으로 네트워크 파라미터 수를 줄여 효율성을 증가시킴과 동시에 성능이 향상된 네트워크 구조를 제안하였다. PFPNet을 포함한 최근의 객체 탐지 알고리즘들은 작은 물체에 대한 탐지 성능을 높이기 위해 피라미드 구조의 feature map들을 사용하고 있다. 우리는 squeeze-and-excitation 네트워크 구조를 수정하여 다양한 해상도의 feature를 확장한 피라미드 squeeze-and-excitation block을 제안했다. 또한 PFPNet의 불필요한 부분을 제거하고 transposed-convolution을 사용하여 더 높은 성능을 보였다. VOC2012에 대한 실험결과는 제안한 네트워크가 기존 네트워크와 비교하여 더 좋은 성능을 갖고 효율적이라는 것을 보여준다.

1. 서 론

영상 내에서 물체의 위치를 탐지하고 식별하는 객체 탐지는 영상 처리에서 유용한 기술이다. 딥러닝이 발달함에 따라 객체 탐지 알고리즘도 발달해왔다. Region-based CNN (R-CNN)[1]은 입력 영상에서 selective search 알고리즘을 사용하여 proposal을 추출한 후, CNN을 통해 feature를 추출하고, soft-max 분류기를 통해 객체가 탐지된다. 이후 등장한 multi-task loss를 사용하는 single stage 모델인 Fast R-CNN[2]은 속도 향상을 위해 제안되었다. Faster R-CNN[3]은 기존 Fast R-CNN에 region proposal 네트워크를 추가하여 anchor 개념을 사용하여 성능을 향상시켰다. 이 후, 객체 탐지 분야는 새로운 구조를 가진 다양한 모델이 개발되어 왔다.

객체 탐지 알고리즘은 크게 두 갈래로 개발이 이루어졌다. 하나는 모델의 속도를 향상시키는 것보다 성능을 향상시키는 것이다. 이러한 예로는 Mask RCNN[4]과 PFPNet[5]이 있다. 다른 하나는 모델의 속도를 올려 효율성을 향상시키는 연구이다. 그 예로 YOLO[6]와 SSD[7]가 있다. 본 논문에서는 더 적은 모델 파라미터를 사용하여 모델의 성능을 향상시키는 네트워크 구조를 제안하였다. 제안하는 네트워크는 PFPNet을 기반으로 한다.

기존 객체 탐지 모델은 작은 물체에 대한 객체 탐지 성능을 높이기 위해 다양한 해상도의 feature map을 사용할 수 있는 피라미드 구조를 이용하였다. 우리는 다양한 해상도의 확장된 feature들을 합치는 과정에서 성능을 올리는 구조를 찾았고, 수정된 SENet[8] 구조를 제안하였다. 제안된 구조는 기존 모델과 비교하여 더 높은 성능을 보였고, 더 효율적임을 확인하였다.

2. 본 론

2.1 Pyramid SE Block

우리는 SE block에서 수정된 새로운 구조의 피라미드 SE block [그림 1]을 제안하였다. SENet [8]의 SE block은 채널 descriptor를 만드는 데 global average pooling을 사용하였지만 이 과정은 너무 간단해서 channel을 대표하는 feature를 충분히 표현하기가 어렵다.

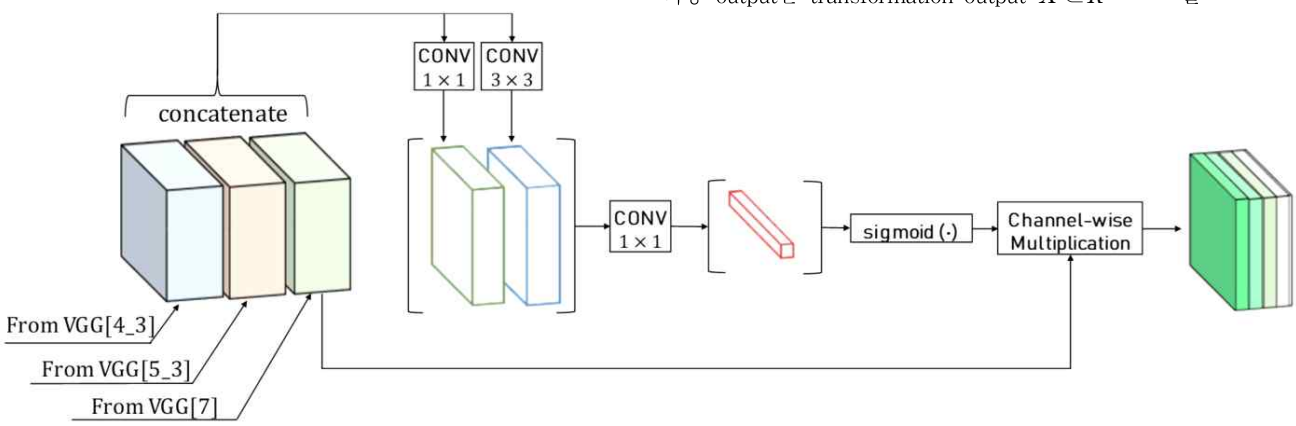
우리는 다양한 크기의 커널들을 사용하여 여러 spatial size에 대해 채널 방향의 feature를 고려하는 convolution을 수행한다. PFPNet에서 PSE block의 입력 feature는 $X \in R^{H \times W \times (C_1 + C_2 + C_3)}$ 이다. C_1, C_2, C_3는 VGG[4_3], VGG[5_3], VGG[7]의 채널 수를 나타낸다. 채널 descriptor는 $z = [z_1, z_2] \in R^{2d}$로 나타낼 수 있다. z는 식 (1)과 같다.

$$z = concat[v_1 * X, v_2 * X] \quad \text{식 (1)}$$

*는 convolution, v_1은 1×1 convolution kernel, v_2는 3×3 convolution kernel이다. 각 convolution에서 추출된 feature 벡터는 병렬로 합쳐지고, 채널의 개수는 d로 조정된다. 이 과정에서, 이후 excitation 파트의 계산량이 줄어든다. SE block은 excitation 파트에서 fully connected layer를 사용한다. 하지만 fully connected layer는 입력 벡터의 채널 수 증가에 비례하여 파라미터수가 증가한다. 제안한 피라미드 SE block은 squeeze 단계에서 채널 수를 조절할 있어 쉽게 파라미터 수를 감소시킬 수 있다. 제안한 block의 excitation 파트의 최종 output s는 다음과 같이 계산된다.

$$s = F_{ex}(z, W) = \sigma(Wz) \quad \text{식 (2)}$$

σ는 sigmoid 함수, $W \in R^{2d \times C_3}$은 컨볼루션 커널이다. block의 최종 output은 transformation output $X' \in R^{H \times W \times C_3}$를



<그림 1> Structure overview of our proposed Pyramid SE block

rescaling해서 얻어진다.

$$\tilde{x_c} = F(x'_c, s_c) = s_c \cdot x'_c \quad \text{식 (3)}$$

$\tilde{x} = [\tilde{x_1}, \tilde{x_2}, \cdots, \tilde{x_{C_3}}]$와 $F(x_C s_C)$는 feature map $x'_C \in R^{H \times W}$와 스칼라 s_C사이의 channel-wise 곱 연산이다. 이 때, 채널 rescaling은 효율적인 계산량을 위해 VGG의 high level feature map인 VGG[7] feature map에 대해서 수행된다.

2.2 Transposed Convolution

PFPNet [5]의 MSCA 모듈의 서로 다른 사이즈의 feature map을 합치기 위해 up-sampling 방법으로 bilienar interpolation을 사용한다. bilinear interpolation은 주로 딥러닝 모델에서 up-sampling을 위해 사용된다. 그러나 pooling으로 down-sampling feature를 증가시킬 때, feature map의 값이 smoothing되어 정보 손실이 일어난다. transposed convoluion은 학습된 파라미터로 입력 feature map의 사이즈를 증가시킬 수 있다.

stride의 값은 up-sampling으로 transposed-convolution을 사용하여 원하는 크기의 output 값을 얻기 위해 조정되어야 한다. stride 값은 파라미터 n이 입력 크기에 의해 나눠진 ouptut 크기 값일 때 다음과 같다.

$$\text{stride}=2n \quad \text{식 (4)}$$

파라미터 수 증가를 막기 위해, 채널 크기 감소를 위한 transposed convolution 이전의, output 채널 크기 16의 1×1 convolution이 진행된다.

2.3 실험결과

정확도를 비교하기 위해 데이터 셋은 <표 1>과 같이 구성하였다.

<표 1> 데이터 셋 구성

Experiment	Dataset
Case 1	VOC 2007(2,000 images)
Case 1+	VOC 2007
Case 2	VOC 2012
Test	VOC 2007 test

Case 1의 base mdel(PFPNet)은 test 데이터 셋에 대해 59.17%의 mAP를 보였고 피라미드 SE block 추가 후 모델은 59.80%으로 향상되었다.<표 2>

<표 2> Case 1에 대한 실험 결과 비교

Model	mAP(%)
PFPNet	59.17
Proposed(w/ PSE)	59.80

Case 1+에서 base model은 64.79% mAP, 피라미드 SE block 추가이후 66.95%로 향상되었다.<표 3>

<표 3> Case 1+에 대한 실험 결과 비교

Model	mAP(%)
PFPNet	64.79
Proposed(w/ PSE)	66.95

Case 2에서 VOC 2007 test datset 4,951 장에 대해 PFPNet은 69.69% mAP를 보였다. PFPNet에 피라미드 SENet을 추가했을 때 모델은 70.85%의 성능을 보였다.<표 4>

<표 4> Case 2에 대한 실험 결과 비교

Model	mAP(%)
PFPNet	69.69
Proposed(w/ PSE+tconv)	70.85

추가로 피라미드 SENet을 추가했을 때 파라미터의 수가 감소하는 것을 확인하였다<표 5>.

<표 5> 파라미터 수 비교

Model	Number of param.	(%)
PFPNet	40,543,512	-
PFPNet w/PSE	34,579,636	14.7↓
PFPNet w/PSE, tconv	34,900,669	13.9↓

3. 결 론

본 논문에서 우리는 더 적은 파라미터로 베이스 모델의 성능을 향상시켰다. 서로 다른 feature를 간단히 조합한 베이스 모델과 비교하여, 우리는 피라미드 SE block이라고 부르는 새로운 feature 합성 방식을 제안하였다. feature 합성 방식은 효과적으로 convolution을 통해 채널 수를 조정함으로써 파라미터 수를 감소시켜 계산량을 줄일 수 있다. 더욱이, 이 방법을 사용하여 feature 합성을 했을 때, 적절한 채널로부터 정보를 얻기 위해 channel attention이 사용되었다. 이것은 또한 객체 탐지의 성능을 높여준다. 제안한 feature 합성 방식은 다른 모델에 적용가능하고 적용했을 때 성능 향상이 기대된다.

[참 고 문 헌]

[1] Girshick, R., Donahue, J., Darrell, T., & Malik, J. "Rich feature hierarchies for accurate object detection and semantic segmentation", In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587, 2014

[2] Girshick, R., ""Fast r-cnn", In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448, 2015

[3] Ren, S., He, K., Girshick, R., & Sun, J. , "Faster r-cnn: Towards real-time object detection with region proposal networks", IEEE transactions on pattern analysis and machine intelligence, 1137-1149, 2016

[4] He, K., Gkioxari, G., Dollár, P., & Girshick, R., "Mask r-cnn", In Proceedings of the IEEE international conference on computer vision, pp. 2961-2969, 2017

[5] Kim, S. W., Kook, H. K., Sun, J. Y., Kang, M. C., & Ko, S. J., "Parallel feature pyramid network for object detection", In Proceedings of the European Conference on Computer Vision (ECCV), pp. 234-250, 2018

[6] Redmon, J., & Farhadi, A., "Yolov3: An incremental improvement", arXiv preprint arXiv:1804.02767., 2018

[7] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C., "Ssd: Single shot multibox detector", In European conference on computer vision, pp. 21-37, 2016

[8] Hu, J., Shen, L., & Sun, G., "Squeeze-and-excitation networks", In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132-7141, 2018