SSD 기반의 잔차 학습 신경망을 이용한 얼굴 검출

*이석희 **장영균 ***조남익

서울대학교

*seokheel@ispl.snu.ac.kr **kyun0914@ispl.snu.ac.kr ***nicho@snu.ac.kr

SSD Based Face Detection using Residual Connections

*Lee, Seok Hee **Jang, Young Kyun ***Cho, Nam Ik

Seoul National University

요약

본 논문은 합성곱 기반의 Single Shot Multibox Detector(SSD) [1] 의 구조를 이용하여 다양한 스케일의 얼굴들을 잘 검출하도록 하였다. 얼굴 검출은 물체 검출과는 다르게 얼굴의 높이와 너비의 비율이 다소 일정하고 크기가 작은 경우가 많은데, 이에 맞게 얼굴 검출이 용이하도록 anchor의 스케일, 비율, 크기를 변경하였다. 특징점 추출 네트워크는 깊은 네트워크의 최적화를 용이하게 하는 skip connection을 이용한 ResNet-50 [2] 기반을 사용하였다. 다양한 크기, 조명, 환경, 각도의 얼굴들을 포함하는 영상들로 이뤄진 Wider Face[3] 데이터 셋의 easy validation set으로 실험한 결과 0.782과 hard validation set에서 0.611의 average precision을 보였다.

1. 서론

얼굴 검출은 얼굴과 관련된 영상 처리 분야 중 가장 기본적이고 중요한 분야이다. 얼굴과 연관된 분야 대부분이 얼굴의 위치를 찾았거나 배경과 분리된 얼굴을 가정으로 시작하기 때문에 정확한 얼굴 검출이 중요하다. 검출이 잘못되었거나 얼굴을 찾지 못했을 경우 해당 어플리케이션의 성능을 크게 떨어뜨릴 수 있기때문에 이상적인 얼굴 검출기는 영상에 존재하는 얼굴들을 모두 검출해야 하고 위치까지도 정확히 추정해야 한다.

얼굴 검출의 고전적인 방법 중 하나인 Viola-Jones[4] 검출기는 hand-crafted 특징점을 추출하고 adaboost 방식을 이용하여 얼굴을 검출 하였다. 고전적인 방법보다 합성곱 신경망을 이용한 딥 러닝 방법이 영상 분류에 뛰어난 성능을 보이면서, 물체 검출 분야에서도 합성곱 신경망을 이용하기 시작했다. 합성곱 신경망을 이용한 물체 검출 알고리즘은 region proposal들을 뽑고 proposal들의 분류와 위치 추정을 하는 방법과 anchor를 기반으로 하는 방법으로 크게 두 가지로 나눌 수 있다.

R-CNN 계열의 검출기는 region proposal들을 뽑는 대표적인 알고리즘들이다. Fast R-CNN[5]은 selective search 방법이나 외부의 region proposal 알고리즘으로 region proposal들을 뽑아냈는데, 수행시간이 오래 걸린다는 단점이 있었다. 이와 달리 Faster R-CNN[6]은 합성곱 신경망으로 이뤄진 region proposal network로 대체해 수행시간을 줄였다.

YOLO[7]는 R-CNN 계열처럼 각 region proposal 마다 분류와 회귀 분석하는 것과 다르게 anchor를 이용하여 입력 영상으로부터 바로 물체를 검출 한다. anchor는 물체 검출을 도와주는 미리 설계된 박스인데, 물체의 크기와 위치는 anchor라는 박스를 기준으로 상대적인 위치와 크기로 표현할 수 있다. YOLO[7]는 다양한 비율과 크기의 anchor를 사용하여 물체의 크기와 비율에도 강인한 검출이 가능하도록 설계하였다.

2. SSD 기반의 얼굴 검출기

Single shot multibox detector (SSD) [1]는 YOLO [7] 구조를 발전시켜 다양한 스케일 물체의 검출에 더 강인하도록 설계를 하였다. YOLO [7] 의 경우 마지막 특징점 맵을 사용하여 물체를 검출했던 반면에, SSD [1]는 마지막 특징점 맵 뿐만 아니라 5개의 중간 특징점 맵을 추가로 사용해 검출을 수행했다. 깊이가 얕은 특징점 맵은 receptive field가 작은 만큼 작은 물체의 검출에 이용하고, 깊은 특징점 맵일수록 보다 큰 물체를 검출 하는데 이용했다.

본 논문의 얼굴 검출기는 학습 시 640 x 640 영상을 입력으로 받는다. 출력은 기존 SSD[1]처럼 6개의 특징점 맵으로부터 해당 anchor에 얼굴이 있을 확률을 나타내는 confidence score와 얼굴이 있을 시 anchor 박스로부터의 상대적 위치를 나타내는 4개 값이다. 기존 SSD[1]의 특징점 추출 네트워크는 VGG 16[8] 기반으로 되어있는데, skip connection을 사용하는 ResNet-50[2]으로 변경하였고, stride가 2인 residual block 2개를 뒤에 추가하였다. 학습 시 대부분의 뉴런이 활성화가 되지 않는 “dead ReLU” 현상이 발생하였고, 이를 피하기 위해 ReLU 부분을 Leaky ReLU로 변경하였다.

물체 검출과 달리 얼굴 검출은 타겟의 크기가 작은 경우가 더 많고 bounding box의 높이와 너비의 비율이 다소 일정하다는 특징이 있다. 이에 따라 <표 1>와 같이 anchor 구조를 변경하였다. Bounding box의 비율이 물체 검출보다 다양하지 않으므로, anchor 비율은 1:1의 정사각형 박스 1 가지만 검출에 사용하였다. 작은 얼굴 검출을 위해 anchor의 stride도 기존과 달리 낮춰, 더 작은 receptive field를 갖는 특징점 맵으로부터 추출하였다. Anchor scale은 stride의 4배로 설정하였다.

Positiion	Stride	Scale	Number
Conv2_3	4	16 x 16	25,600
Conv3_3	8	32 x 32	6,400
Conv4_3	16	64 x 64	1,600
Conv5_3	32	128 x 128	400
Conv6_3	64	256 x 256	100
Conv7_3	128	512 x 512	25

<표 1> Anchor 구조

일반적으로 얼굴 검출의 sample들은 negative 샘플들이 대부분의 비율을 차지하고 있다. Data driven 방법 중 하나인 딥 러닝 네트워크를 imbalance한 데이터 그대로 학습하게 되면 검출기의 결과 대부분을 false positive로 추론하기 때문에 성능 저하를 가져올 수 있다. 따라서 분류 목적 함수는 <식 1>처럼 Focal loss[9]을 사용하였다. Focal loss[9]는 easy sample들의 loss를 줄이고 hard sample들의 학습에 집중하여 앞서 말한 문제를 감소시킨다고 알려져 있다. N_p와 N_n는 각각 positive와 negative anchor의 개수를 나타내고, p는 positive 혹은 negative anchor의 confidence score이다. <식 1>에서의 r은 1을 사용하였다. Regression 목적 함수는 일반적으로 많이 사용하는 smooth L1 loss를 사용하였다.

$$L_c = \frac{1}{N_p}\left(\sum_{i}^{N_p} -(1-p_i)^r \log(p_i) + \sum_{k}^{N_n} -(1-p_k)^r \log(p_k)\right)$$

<식 1> 분류 목적 함수로 사용된 Focal Loss

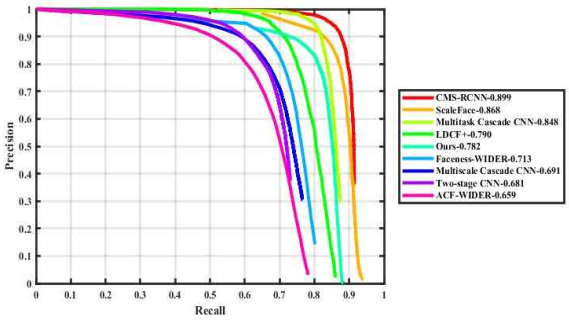
3. 실험

네트워크 학습은 Wider Face[3]의 training set으로 학습하였다. Batch normalization의 momentum은 0.9로 설정하였다. Batch 크기는 6으로 설정하였고, momentum optimizer를 사용하여 learning rate를 1e-3부터 시작해서 50,000 step 마다 1/10 만큼 감소 시켜 200,000 step 까지 학습했다.

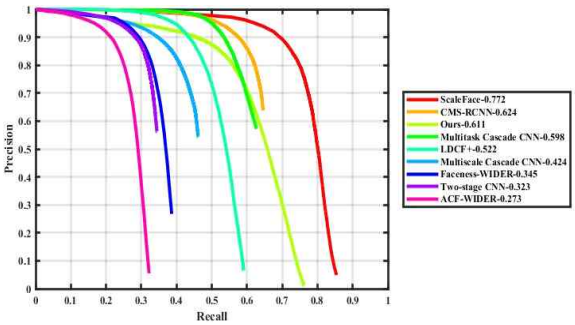
Anchor와 ground truth 박스의 intersection over union (IoU)이 0.35 이상이면 positive anchor로 설정하였고, 0.3 미만이면 negative anchor로 설정하였다. 그리고 앞 섹션에서 언급했듯이 대부분의 anchor가 negative anchor가 되어 false positive rate이 올라가는 것을 방지하기 위하여, 일반적으로 많이 사용하는 hard sample mining 기법을 이용하였다. 따라서 실제 학습에 사용하는 negative와 positive sample의 비율을 3:1로 맞췄다.

4. 실험 결과

Wider Face[3]의 validation set에 대해 <그림 2, 3>와 같이 precision-recall curve를 그려 성능을 평가하였다. Wider Face 데이터셋의 경우 얼굴의 크기와 난이도에 따라 easy, medium, hard로 나눠진다. 실험 결과 easy 셋에 대해 average precision (AP)은 0.782로 [10]과 비슷한 성능을 보였다. Hard 셋에 대해 AP값은 0.611이었고, 작은 얼굴에 대한 검출율은 easy셋 만큼 미치지 못하지만 [10]보다 더 잘 검출했다는 것을 알 수 있다. <그림 4>과 같이 비교적 큰 얼굴들은 높은 confidence score로 검출이 가능했지만 작은 얼굴의 경우 검출을 하지 못하거나 false positive가 발생하는 것을 확인했다.



<그림 2> Wider Face easy validation에 대한 precision-recall



<그림 3> Wider Face hard validation에 대한 precision-recall

<그림 4> 얼굴 검출 결과

5. 결론

본 논문에서는 잔차 연결을 이용한 신경망을 이용한 SSD를 얼굴 특성에 맞게 네트워크와 anchor 구조를 변경하였다. Hard negative sample을 학습하는 것을 집중하기 위해 hard sample mining과 focal loss를 적용하였다. Wider Face 데이터셋의 easy validation set에서 실험한 결과 0.782였고, hard validation set에서 0.611이었다. 앞으로 연구는 더 다양한 특징점 맵에서 얼굴 검출을 수행하여 검출율을 높이거나, anchor 구조를 변경하여 작은 얼굴과 매칭되는 positive anchor의 숫자를 높히는 방향으로 진행할 계획이다.

감사의 글

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No. 1711075689, AI 어플리케이션을 지원하는 IoT 연동 분산 엣지-클라우드 기술 개발)

참고문헌

[1] Liu, Wei, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. "Ssd: Single shot multibox detector." In European conference on

computer vision, pp. 21-37. Springer, Cham, 2016.
[2] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.
[3] Yang, Shuo, Ping Luo, Chen-Change Loy, and Xiaoou Tang. "Wider face: A face detection benchmark." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5525-5533. 2016.
[4] Viola, Paul, and Michael Jones. "Rapid object detection using a boosted cascade of simple features." CVPR (1) 1 (2001): 511-518.
[5] Girshick, Ross. "Fast r-cnn." In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448. 2015.
[6] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." In Advances in neural information processing systems, pp. 91-99. 2015.
[7] Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788. 2016.
[8] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
[9] Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. "Focal loss for dense object detection." In Proceedings of the IEEE international conference on computer vision, pp. 2980-2988. 2017.
[10] Ohn-Bar, Eshed, and Mohan M. Trivedi. "To boost or not to boost? on the limits of boosted trees for object detection." In 2016 23rd international conference on pattern recognition (ICPR), pp. 3350-3355. IEEE, 2016.