Libra R-CNN: Towards Balanced Learning for Object Detection

CVPR 2019

Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, Dahua Lin

Zhejiang University, The Chinese University of Hong Kong, Sense Time Research, The University of Sydney

Abstract

object detection 문제를 다룰 때, detection에 대한 문제가 상대적으로 덜 주목 받고 있었다.

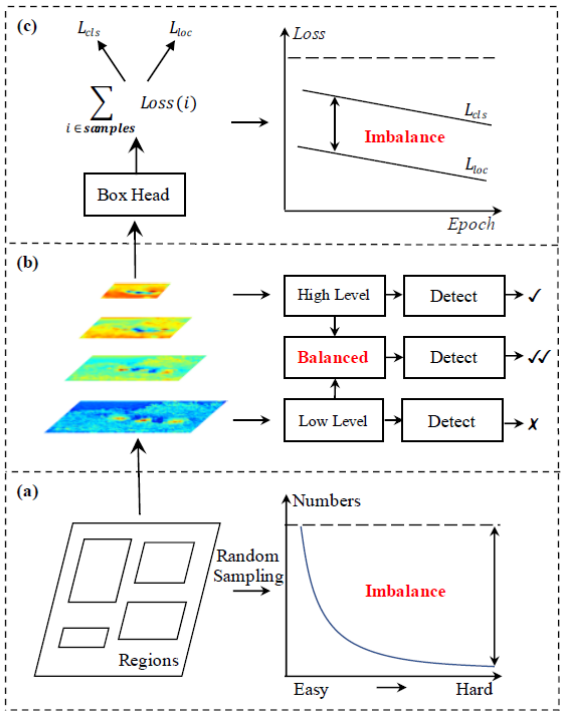
training process는 sample level, feature level, object level로 구성되어 있는데, 이 과정에 있어서 불균형으로 인한 성능제한 문제점이 발생한다.

Libra R-CNN 은 간단하지만, 객체검출에 대하여 균형적인 학습(balanced learning)에 대하여 효과적이다.

Libra R-CNN은 balanced learning을 위하여 IoU-balanced sampling, balanced feature pyramid, balanced L1 loss을 통해 기존 불균형 문제점을 개선하였다.

결과적으로 FPN Faster R-CNN과 RetinaNet보다 AP성능에 대하여 MS COCO에서 2.5 points, 2.0 points 더 높게 나타났다.

Introduction



Object detection는 region을 sampling 하고 거기서 feature를 추출하고 recognizing the categories와 refining location을 하는 공통적인 pipeline 패러다임을 가지고 있다.

따라서, object detection의 training 성공은 다음 3가지의 측면에 의존된다.

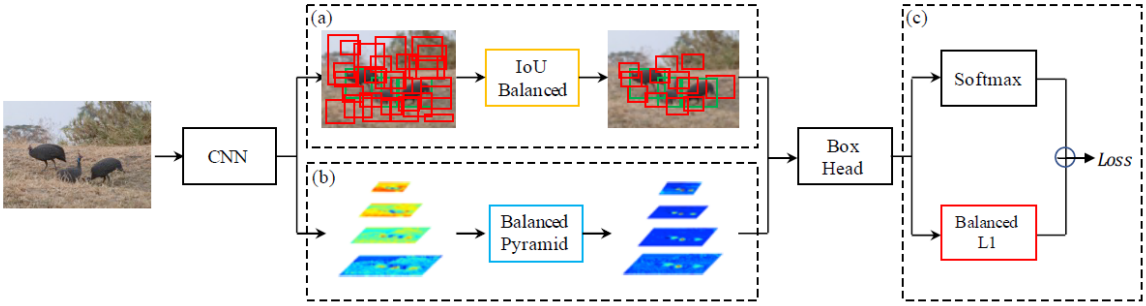
1. whether the selected region samples are representative
2. whether the extracted visual features are fully utilized
3. whether the designed objective function is optimal

하지만, 위 3가지 측면 모두 상당히 불균형하고 이는 기존 잘 설계된 모델 구조의 성능을 제한한다.

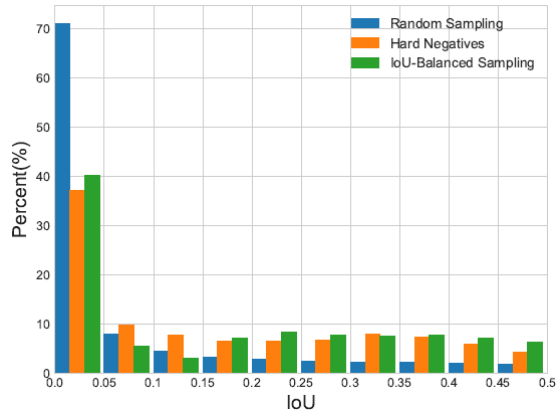
Libra R-CNN

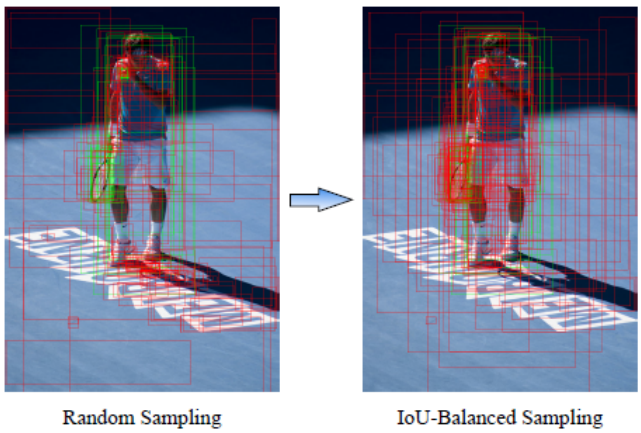
- Goal

> *전체적인 균형 잡힌 설계를 사용하여 detector의 training process에서 존재하는 불균형을 완화하여 모델 아키텍쳐의 잠재력을 이용하는것이다.*



- IoU-balanced Sampling





Random Sampling IoU-Balanced Sampling

Settings	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline	35.9	58.0	38.4	21.2	39.5	46.4
Pos Balance	36.1	58.2	38.2	21.3	40.2	47.3
$K = 2$	36.7	57.8	39.9	20.5	39.9	48.9
$K = 3$	36.8	57.9	39.8	21.4	39.9	48.7
$K = 5$	36.7	57.7	39.9	19.9	40.1	48.7

$$p = \frac{N}{M} \tag{1}$$

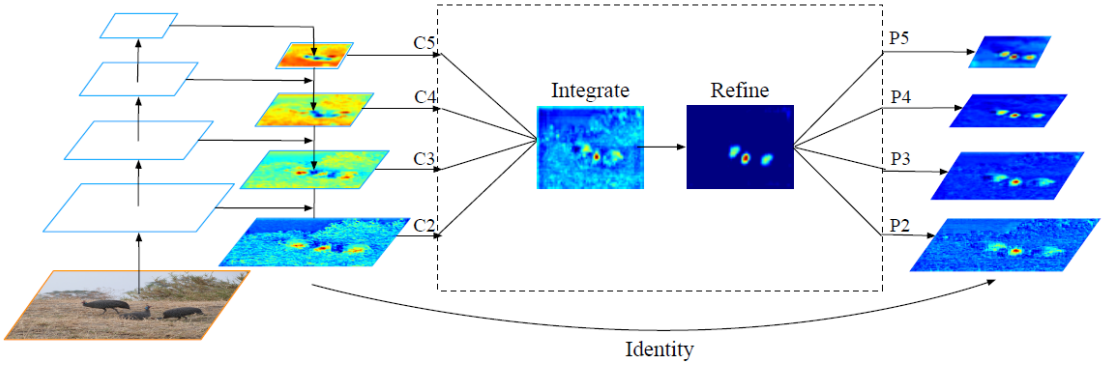
$$p_k = \frac{N}{K} * \frac{1}{M_k}, \quad k \in [0, K) \tag{2}$$

- N : negative samples
- M : corresponding candidates
- p : probability for each sample under random sampling
- K : IoU에 따른 샘플링 간격(default=3)
- M_k : number of sampling candidates in the corresponding interval denoted by k

> -. hard negative sample에 대해 중점적으로 고려하여 balance sampling을 진행한다.
>
> -. hard negative는 IoU가 0.05보다 큰 경우가 60%를 차지하지만, Random Sampling의 경우, 30%정도만 제공되고 있는것을 알 수 있다.
>
> -. IoU balanced sampling을 통해 hard negative sample에 대해 filtering 할 수 있다.
>
> -. IoU가 높은 샘플이 선택될 가능성이 높을수록 K에 민감하지 않다.

- Balanced Feature Pyramid

> *key idea*
>
> *같은 깊이의 통합된 balanced semantic features를 사용하여 multi-level features를 강화하는 것.*



  - Obtaining balanced semantic features

    각 resolution에 대한 semantic feature 추출 -> 각 features를 하나의 size(ex. $C4$)로 만들어 줌(interpolation, max-pooling) -> Averaging

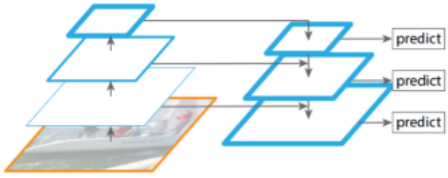
$$C = \frac{1}{L} \sum_{l=l_{min}}^{l=l_{max}} C_l. \tag{3}$$

Note that this procedure does not contain any parameter

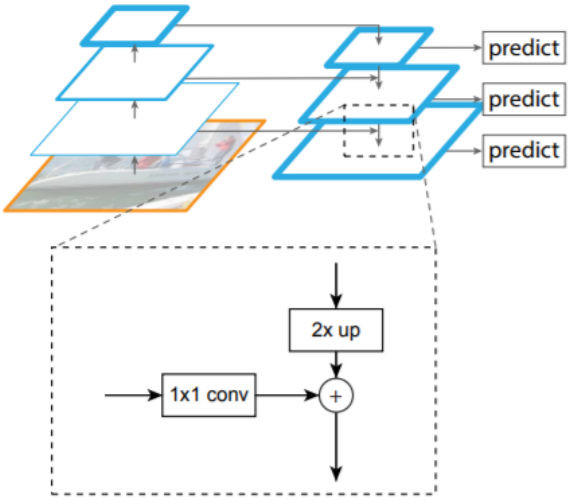
- Refining balanced semantic features

  본 논문에서는 기본적으로 Gaussian을 활용한 non-local attention을 사용하였으며, output인 {P2, p3, p4, p5}는 FPN과 같은 pipeline을 적용하였다.

> FPN(Feature Pyramid Network) pipeline



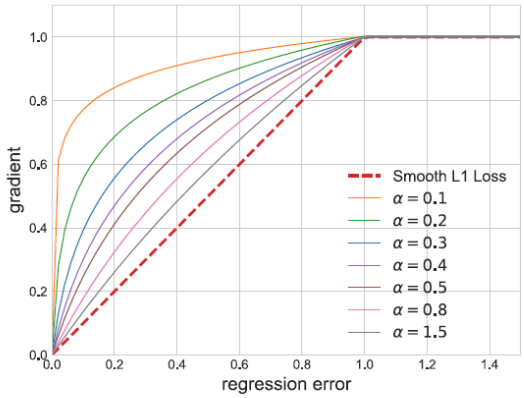
(d) Feature Pyramid Network



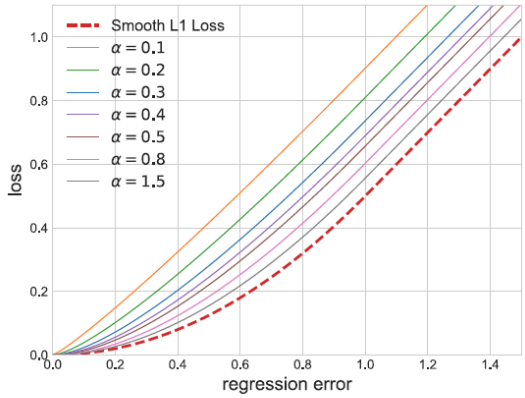
- Balanced L1 Loss (L_b)

$$L_{p,u,t^u,v} = L_c ls(p,u) + \lambda[u \geq 1] L_l oc(t^u, v) \quad (4)$$

  - inlier는 outlier와 비교했을 때, 전체의 샘플 당 gradients 평균의 30%밖에 차지하지 않는다. (loss >= 1.0 : outlier / others : inlier)



(a)



(b)

> *key idea*
>
> Smooth L1 Loss에서 파생되었으며, outlier가 생성하는 변곡점부분을 clipping한다.
>
> 결정적인 regression gradients를 promoting 하는 것이다. 즉, inlier로 부터의 gradients를 promoting하여 관련 샘플과 작업의 균형을 재조정하는 것이다.

- Localization loss defined as

$$L_{loc} = \sum_{i \in x,y,w,h} L_b(t_i^u - v_i) \tag{5}$$

- corresponding formulation of gradients follows

$$\frac{\partial L_{loc}}{\partial w} \propto \frac{\partial L_b}{\partial t_i^u} \propto \frac{\partial L_b}{\partial x} \tag{6}$$

- based on (6), promoted gradient formulation as

$$\frac{\partial L_b}{\partial x} = \begin{cases} \alpha \ln(b|x|+1) & \text{if } |x| < 1 \\ \gamma & \text{otherwise} \end{cases} \tag{7}$$

- balanced L1 loss

$$L_b = \begin{cases} \frac{\alpha}{b}(b|x|+1)\ln(b|x|+1) - \alpha|x| & \text{if } |x| < 1 \\ \gamma|x| + C & \text{otherwise} \end{cases} \tag{8}$$

- parameters γ, α, and b

$$\alpha \ln(b+1) = \gamma \tag{9}$$

- default parameters : α = 0.5 / γ = 1.5

Settings	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline	35.9	58.0	38.4	21.2	39.5	46.4
loss weight = 1.5	36.4	58.0	39.7	20.8	39.9	47.5
loss weight = 2.0	36.2	57.3	39.5	20.2	40.0	47.5
L1 Loss (1.0)	36.4	57.4	39.1	21.0	39.7	47.9
L1 Loss (1.5)	36.6	57.2	39.8	20.2	40.0	48.2
L1 Loss (2.0)	36.4	56.5	39.6	20.1	39.8	48.2
$\alpha = 0.2, \gamma = 1.0$	36.7	58.1	39.5	21.4	40.4	47.4
$\alpha = 0.3, \gamma = 1.0$	36.5	58.2	39.2	21.6	40.2	47.2
$\alpha = 0.5, \gamma = 1.0$	36.5	58.2	39.2	21.5	39.9	47.2
$\alpha = 0.5, \gamma = 1.5$	37.2	58.0	40.0	21.3	40.9	47.9
$\alpha = 0.5, \gamma = 2.0$	37.0	58.0	40.0	21.2	40.8	47.6

Experiments / Results

- 모든 experiments는 MS COCO dataset을 활용하였으며, COCO test-dev에서 2019 SOTA object detection approach 달성

Method	Backbone	Schedule	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
YOLOv2 [27]	DarkNet-19	-	21.6	44.0	19.2	5.0	22.4	35.5
SSD512 [23]	ResNet-101	-	31.2	50.4	33.3	10.2	34.5	49.8
RetinaNet [20]	ResNet-101-FPN	-	39.1	59.1	42.3	21.8	42.7	50.2
Faster R-CNN [19]	ResNet-101-FPN	-	36.2	59.1	39.0	18.2	39.0	48.2
Deformable R-FCN [6]	Inception-ResNet-v2	-	37.5	58.0	40.8	19.4	40.1	52.5
Mask R-CNN [10]	ResNet-101-FPN	-	38.2	60.3	41.7	20.1	41.1	50.2
Faster R-CNN*	ResNet-50-FPN	1×	36.2	58.5	38.9	21.0	38.9	45.3
Faster R-CNN*	ResNet-101-FPN	1×	38.8	60.9	42.1	22.6	42.4	48.5
Faster R-CNN*	ResNet-101-FPN	2×	39.7	61.3	43.4	22.1	43.1	50.3
Faster R-CNN*	ResNeXt-101-FPN	1×	41.9	63.9	45.9	25.0	45.3	52.3
RetinaNet*	ResNet-50-FPN	1×	35.8	55.3	38.6	20.0	39.0	45.1
Libra R-CNN (ours)	ResNet-50-FPN	1×	38.7	59.9	42.0	22.5	41.1	48.7
Libra R-CNN (ours)	ResNet-101-FPN	1×	40.3	61.3	43.9	22.9	43.1	51.0
Libra R-CNN (ours)	ResNet-101-FPN	2×	41.1	62.1	44.7	23.4	43.7	52.5
Libra R-CNN (ours)	ResNeXt-101-FPN	1×	43.0	64.0	47.0	25.3	45.6	54.6
Libra RetinaNet (ours)	ResNet-50-FPN	1×	37.8	56.9	40.5	21.2	40.9	47.7

Conclusion

본 논문에서는 detector의 training process에 있어서 본래의 설계된 구조의 성능에 비하여 potential 이 잘 나타나지 않고 있었음을 말하고 있다.

이에 Libra R-CNN을 제안하며, training process에 있어서 발생된 imbalanced를 해결하기위해, IoU-balanced sampling, balanced feature pyramid, balanced L1 loss 기법을 적용하였다.

MS COCO dataset을 활용하여 간단하지만 매우 효과적인 성능을 발휘하였으며, single stage / two stage detector 모두 backbone 사용에 있어 일반화가 잘 되어 있다.