Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun
Summarized by 202355514 강지원

1 Faster R-CNN

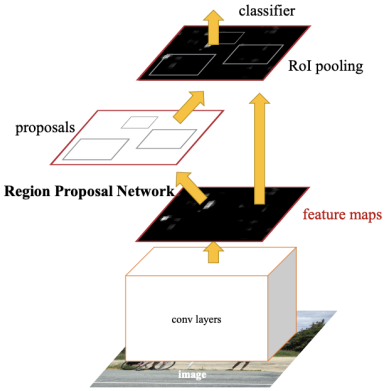

Figure 1: Faster R-CNN

Faster R-CNN은 Region Proposal Network(RPN)과 Fast R-CNN으로 구성된 단일 네트워크로, 두 모듈이 합성곱 계층(Feature Map)을 공유하며 연산의 효율성을 높인다. RPN은 입력 이미지로부터 후보 영역과 객체성 점수를 직접 예측하는 완전 합성곱 신경망으로, 슬라이딩 윈도우 방식으로 특징을 추출한 후, 박스 회귀와 분류를 동시에 수행한다.

1.1 Anchor

RPN에서는 각 슬라이딩 윈도우 위치마다 다수의 후보 영역을 동시에 예측하며, 이를 위해 Ancho라 불리는 기준 박스를 도입한다. 기존 방식인 MultiBox는 많은 수의 anchor(800개)를 만들어 놓고, 위치에 따라 다 다르게 학습해야 해서 모델 크기도 크고 위치 이동에 약했다. 반면 Faster R-CNN의 Anchor 방식은 단순히 각 위치마다 9개 고정된 기준 박스를 두고, 그 박스를 기준으로 후보 박스를 회귀하는 방식이다. 따라서 모델이 훨씬 단순해지고, 위치 변화에도 강건하며, 본 연구에서 제안하는 Anchor 기반 다중 스케일 설계는 미리 정의된 다양한 크기의 Anchor를 기준으로 bounding box를 분류 및 회귀하는 방식을 취한다. 이 접근법은 단일 스케일의 이미지와 단일 크기의 필터만을 사용하면서도 다중 스케일 문제를 효과적으로 해결한다. 실험 결과, 본 방식은 추가 연산 비용 없이 합성곱 특징을 공유할 수 있는 효율적인 다중 스케일 처리 방안임을 확인하였다.

1.2 Loss

RPN의 학습은 Anchor에 대해 객체와 비객체로 분류하는 이진 분류와 Bounding Box 회귀를 동시에 수행하는 다중 과제 손실(multi-task loss)로 정의된다. 우선 IoU란 Intersection over Union의 약자로, 겹치는 영역의 크기를 두 영역의 총합으로 나눈 값이다. 이는 0에서 1 사이의 값을 가지며 1에 가까울수록 두 개의 바운딩 박스가 겹치는 영역이 크고 모델 출력값이 정확하다는 것을 의미한다. Anchor는 IoU $\geq$ 0.7이거나 Ground Truth 및 최대 IoU를 갖는 경우 양성으로, IoU $\leq$ 0.3이면 음성으로 지정된다.

1.3 Training RPNs

RPN은 역전파와 SGD를 통해 end-to-end 방식으로 학습된다. 미니배치는 단일 이미지에서 256개의 Anchor를 샘플링하여 구성하며, 양성과 음성 Anchor의 비율은 최대 1:1로 유지된다. 새로운 계층은 Gaussian 분포로 초기화하고, 공유 합성곱 계층은 ImageNet 분류 모델로 사전 학습된 가중치를 사용한다. ZF 모델은 전체 계층을, VGG-16 모델은 conv3_1 이후 계층을 fine-tuning하였으며, 학습률은 0.001에서 시작하여 후반부에 0.0001로 조정하였다.

2 Sharing Features for RPN and Fast R-CNN

RPN과 Fast R-CNN은 convolutional layer를 공유하는 단일 네트워크로 학습될 수 있다. 이를 위해 세 가지 학습 전략이 있다. 첫째, Alternating Training(교대 학습)은 RPN과 Fast R-CNN을 번갈아 가며 학습시키는 방식으로 본 논문의 실험에서 사용되었다. 둘째, Approximate Joint Training(근사적 공동 학습)은 두 네트워크를 통합하여 역전파 시 손실을 함께 전달하는 방식으로, proposal 좌표에 대한 gradient는 무시되지만 학습 시간이 크게 단축된다. 셋째, Non-Approximate Joint Training(정확한 공동 학습)은 proposal 좌표에 대한 gradient까지 고려해 구현이 복잡하다. 본 연구에서는 실제로 4 -Step-Alternating Training 절차를 적용하여 두 네트워크의 convolutional layer를 공유하는 통합 모델을 구축하였다.

3 Experiments

3.1 PASCAL VOC

PASCAL VOC 2007에서의 평가 결과, RPN과 Fast R-CNN을 결합한 모델은 Selective Search(58.7%) 및 Edge Boxes(58.6%)와 비교해 경쟁력 있는 성능을 보이며, 300개 proposal만으로도 59.9%의 mAP를 달성하였다. 이는 convolutional layer를 공유함으로써 제안 수를 줄이고도 정확도를 유지할 수 있음을 보여준다. cls 점수는 상위 proposal의 정확도를, reg는 박스 품질 향상에 기여하며, VGG-16 기반 RPN은 ZF 기반보다 더 높은 proposal 품질을 달성하였다. VGG-16 기반 RPN은 이는 Selective Search보다 빠른 결과를 보였다.
OverFeat와 같은 One-Stage 탐지기는 슬라이딩 윈도우 기반으로 위치와 클래스를 동시에 예측하지만, 연산량이 많고 정확도가 낮다. 반면 Faster R-CNN은구성된 Two-Stage 구조를 취한다. PASCAL VOC 실험에서 Two-Stage 방식은 58.7%로 약 4.8%p 향상되었다. 이는 계단식(cascade) 구조의 Region Proposal과 탐지 단계가 객체 탐지의 정확도와 효율성을 모두 향상시킴을 보여준다.

3.2 MS COCO

MS COCO 실험에서 Faster R-CNN은 Fast R-CNN 대비 mAP@0.5 기준 2.8%, mAP@[.5:.95] 기준 2.2% 향상된 성능을 보였다. ResNet-101과 같은 심층 네트워크를 도입하면 성능이 크게 향상되었다. 이는 RPN이 학습 기반의 region proposal로서 깊은 특성과 결합하여 객체 탐지 정확도를 크게 높일 수 있음을 입증한다.

3.3 From MS COCO to PASCAL VOC

대규모 데이터셋의 활용 효과를 검증하기 위해 MS COCO에서 학습된 모델을 PASCAL VOC에 적용하였다. COCO 모델을 VOC 데이터에 직접 적용한 경우에도 76.1% mAP을 달성하여 VOC 데이터만으로 학습한 모델(73.2%)보다 우수한 성능을 보였다. 따라서 MS COCO에서 학습된 모델을 VOC에 적용하거나 fine-tuning하면 성능이 크게 향상됨으로써 대규모 데이터셋의 가치를 입증하였다.

4 Conclusion

본 연구에서는 효율적이고 정확한 Region Proposal 생성을 위해 RPN을 제안하였다. 제안된 RPN은 탐지 네트워크와 합성곱 특징을 공유함으로써 제안 단계의 연산 비용을 최소화하며, 이를 통해 통합된 객체 탐지 시스템을 실시간에 가까운 속도로 구현할 수 있다. 또한 학습된 RPN은 제안 품질을 개선하여 전체 객체 탐지 정확도를 향상시킨다.

[1] IoU 개념 정리: https://silhyeonha-git.tistory.com/3#google_vignette