SSD: Single Shot MultiBox Detector

* 한 개의 DNN을 이용해 이미지에서 객체를 탐지하는 메소드 제시함

* SSD라고 불리는 논문의 접근은 bounding box의 output space를 디폴트 박스들의 세트로 나눔

* 예측 시간에, 네트워크는 각 디폴트 박스에 있는 각 객체 카테고리의 존재에 대한 점수를 만들고, 박스의 수정을 생성해 객체 모양을 더 좋게 짝을 맞춤. 또한, 네트워크는 다양한 사이즈의 객체를 자연스럽게 다루기 위해 다양한 feature map에서 얻은 예측과 다양한 해상도를 조합함

* Object proposal를 요구하는 메소드들에 비교하여 상대적으로 심플한데, 그 이유는 그것은 proposal generation과 연속되는 픽셀 또는 feature resampling stage를 완전히 삭제하고 모든 계산을 하나의 네트워크로 요약하기 때문임

-> 이것은 SSD가 훈련하기 쉽게 만들고 탐지 요소를 요구하는 시스템으로 통합하는 것을 어렵지 않게 만들어 줌

* PASCAL VOC, COCO, and ILSVRC 데이터셋의 실험적 결과는 SSD가 추가적인 객체 제안 단계를 활용하는 메소드들에 비해 경쟁적인 정확도를 가지고 있고, 훈련과 추론 모두를 위한 통일된 프레임워크를 제공하는 동안 훨씬 빠르다는 것을 입증함

* 다른 single stage 메소드들에 비해 훨씬 작은 입력 이미지 사이즈에서도 훨씬 좋은 성능을 가지고 있음

* 탐지 파이프라인의 각 단계를 attack 함으로써 빠른 detector를 만들기 위한 많은 시도가 있었음

-> 그렇지만 지금까지는 매우 향상된 속도는 매우 감소된 탐지 성능으로만 얻을 수 있음

-> 이 논문은 높은 정확도의 탐지를 위한 매우 큰 향상된 결과가 나옴

* 속도에서 가장 근본적인 발전은 bounding box, subsequent pixel, 또는 feature resampling stage를 제안하는 것을 삭제하는 것에서 옴

-> 이 논문에서 처음 하는 것은 아니지만, 발전의 시리즈을 더함으로써, 논문에서는 기존 시도들보다 성능을 매우 향상하는 것을 가능하도록 했음

-> 발전은 다양한 aspect ratio detector를 위해 별개의 predictor을 이용함으로써, 그리고 multiple scale에서 감지를 수행하기 위해 네트워크의 후기 단계에서 온 다양한 feature map을 이러한 필터들을 곱하는 것을 적용함으로써 bounding 박스 위치의 객체 카테고리들과 오프셋들을 예측하기 위해 작은 convolutional filter를 이용하는 것을 포함

* 이 논문의 기여

1) 다양한 카테고리를 위한 Single-shot detector, SSD를 소개함

- single shot detector를 위한 기존 SOTA였던 YOLO보다 빠름

- Faster R-CNN를 포함한 explicit region proposal과 pooling을 수행하는 느린 기술보다 빠름

2) SSD의 핵심은 feature map에 적용한 작은 convolutional 필터를 이용해 default bounding box의

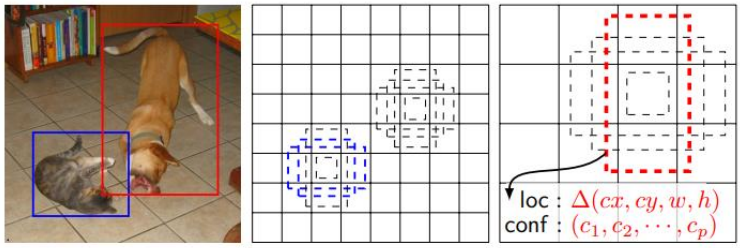
고정된 세트를 위한 카테고리 점수와 box offset를 예측하는 것

3) 높은 탐지 정확도를 얻기 위해 다양한 범위의 feature map에서 다양한 범위의 예측을 생산하고, 측면 범위에서 예측들을 명시적으로 분리함

4) 이러한 디자인 특성들은 심지어 낮은 해상도의 입력 이미지에서도 간단한 end-to-end 훈련과 높은 정확도로 이끌고. 또한 속도 vs 정확도 trade-off를 발전시킴

5) PASCAL, VOC, COCO, ILSVRC에서 평가된 다양한 입력 크기를 가진 모델에서 timing과 정확도 분석을 포함한 실험들을 했을 때 최근 SOTA 접근들과 비교됨

* Framework



(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

- SSD는 훈련 동안 각 객체를 위해 오직 input image와 ground truth box들을 필요로 함

* Model

- feed-forward convolutional network -> bounding box의 fixed-size collection과 그러한 box의 객체 class instance의 존재에 대한 점수를 만듦. 마지막 탐지를 하기 위해 non-maximum suppression 단계가 뒤이어 나옴

- 초기 네트워크 레이어는 표준 아키텍처를 기반으로 하고 있음. 그 다음 네트워크에 보조의 구조를 추가하여 탐지를 함. 다음과 같은 주요 특성을 가지고 있음

Multi-scale feature map for detection

-> Truncated base network의 end에 convolutional feature layer들을 추가함. 이러한 레이어들은 계속해서 사이즈가 줄어들고, 다양한 범위의 탐지 예측을 가능하게 함

Convolutional predictors for prediction

-> 각각 더해진 feature 레이어는 convolutional feature 세트를 사용하여 탐지 예측의 고정된 세트를 만듦

Default boxes and aspect ratios

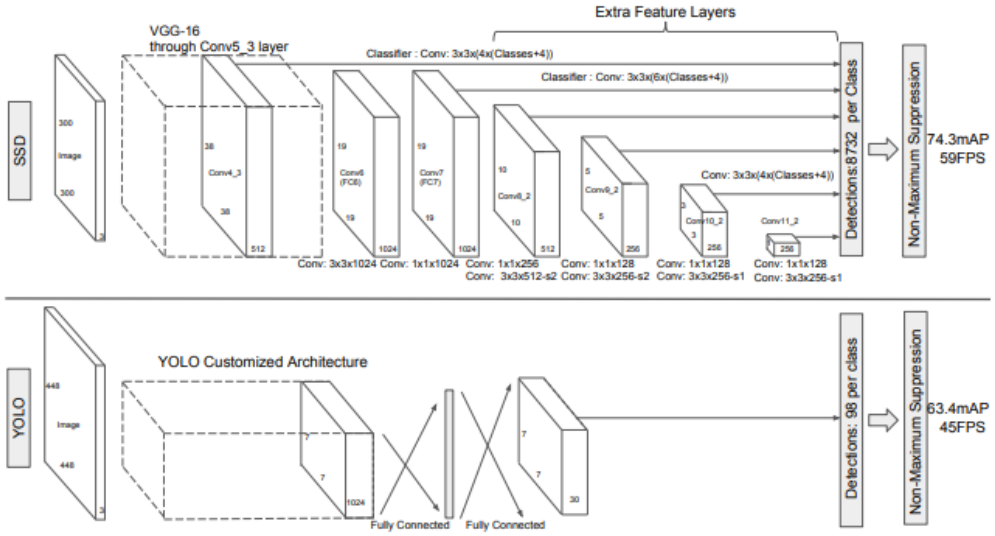
-> Default bounding box set을 각각의 feature map cell과 연관 지음. Default box는 feature map을 convolutional 면에서 타일링함(그로 인해 연관된 cell과 상대적인 각 box의 위치가 고정될 수 있음)

-> Default box는 Faster R-CNN에서 사용된 anchor box들과 유사하지만, 다른 해상도를 다양한 feature map에 적용함

-> 다양한 feature map에서 다른 default box shape를 허용하는 것은 가능한 출력 box shape의 공간을 효율적으로 별개의 것으로 구분하게 함

- SSD와 YOLO 차이

: SSD는 base network의 end에 몇몇 feature 레이어를 추가하여 다양한 범위의 default box와 aspect ratio들, 그리고 그들의 연관된 confidence에서 offset을 예측함



* Conclusion

1) 다양한 카테고리를 위한 fast single-shot object detector를 소개함

- SSD의 주요 특성은 네트워크의 top에 있는 다양한 feature map에 부착된 multi-scale convolutional bounding box 출력의 사용임

-> 가능한 box shape 공간을 효율적으로 모델링하게 함

2) 논문의 저자들은 실험적으로 주어진 적절한 훈련 전략, 많은 수의 조심스럽게 주어진 default bounding box들이 더 향상된 성능으로 이어진다는 것을 입증함

3) 기존의 메소드들보다 box prediction sampling location, scale, aspect ratio가 최소 몇 배 이상 높은 SSD 모델을 만듦

4) 같은 VGG-16 베이스 아키텍처가 주어졌을 때, SSD는 기존 SOTA 객체 detector counterpart보다 성능과 속도 면에서 순조롭게 비교함

5) SSD512 모델은 SOTA인 Faster R-CNN보다 PASCAL VOC과 COCO에서 성능을 압도했고, 속도도 3배 빠름

* Real time SSD300 모델은 59FPS으로 실행되는데, 이것은 더 나은 탐지 성능을 만들면서 현재 real time YOLO alternative보다 빠름.

* Standalone utility를 제외하고, 논문의 저자들은 획일적이고 상대적으로 간단한 SSD 모델이 객체 탐지 구성 요소를 이용하는 큰 시스템을 위한 유용한 building block을 제공한다고 믿음

* 비디오의 객체를 동시에 탐지하고 추적하기 위해 recurrent neural network를 사용하여 시스템의 부분으로써 쓰임을 연구하는 것이 미래 계획임