Summary in
“Recent Advances in Deep Learning for Object Detection”

AI Lab
Chang Gi Moon

목차

3. **keypoint 기반 기법들(anchor-free 기법들)**
    1. corner 기반 기법들
    2. center 기반 기법들
4. **기타 기법들(keypoint 또는 anchor에 기반하지 않는 기법들)**

5. **Feature Representation Learning**
    1. **다중-스케일 특징 학습(multi-scale feature learning)**
        1. 이미지 피라미드(Image Pyramid)
        2. 통합 특징(Integrated Features)
        3. 예측 피라미드(Prediction Pyramid)
        4. 특징 피라미드(Feature Pyramid)
    2. **영역 특징 부호화(region feature encoding)**
    3. **문맥 추론(contextual reasoning)**
        1. 전역 문맥 추론(global contextual reasoning)
        2. 영역 문맥 추론(region contextual reasoning)
    4. **변형 가능한 특징 학습(deformable feature learning)**

6. **Learning Strategy**
    1. **Training Stage**
        1. **데이터 증강(data augmentation)**
        2. **불균형 샘플링(imbalance sampling)**
            1. 클래스 불균형(class imbalance)
            2. 난이도 불균형(difficulty imbalance)
        3. **localization 정제**
            1. 기본적인 localization 정제
            2. auxiliary models 학습을 이용한 localization 추가 정제
            3. 통합 프레임워크 설계를 위한 object function 수정

4. **Cascade 학습**
5. **기타 전략(다른 종류의 학습 전략)**
    1. 적대적 학습(adversarial learning)
    2. 밑바닥부터 훈련하기(learning from scratch)
    3. 지식 증류(knowledge distillation)

2. **Test Stage**
    1. **중복 제거(duplicate removal)**
        1. NMS 개선
    2. **모델 가속(model acceleration)**
        1. 계산 비용 공유 또는 feature map 채널 축소
        2. detection backbone을 개선
        3. 모델 압축 및 양자화 이용
    3. **기타 학습 전략(image transformation 적용)**
        1. 이미지 피라미드 또는 horizontal flipping 적용

7. Application

1. **Face Detection**
    1. 다중 스케일 특징 학습(multi-scale feature learning)을 이용한 연구들
    2. 문맥 정보(context information)를 활용한 연구들
    3. 기타 다양한 연구들
2. **Pedestrian Detection**
    1. 기본적인 다양한 연구들
    2. 가려짐(occlusion)을 다루기 위한 연구들
3. **기타 Application**
    1. logo detection
    2. video object detection
    3. vehicle detection, traffic-sign detection, skeleton detection

8. Detection Benchmarks

1. Generic Object Detection Benchmarks
2. Face Detection Benchmarks
3. Pedestrian Detection Benchmarks

9. SOTA for Generic Object Detection

1. Pascal VOC 2007 & VOC 2012 결과
2. MSCOCO 결과

10. Concluding Remarks & Future Directions

1. 규모에 가변적인(scalable) proposal generation 전략
2. 문맥 정보(contextual information)의 효율적인 부호화(encoding)
3. 자동 기계 학습(AutoML)에 기반한 detection
4. object detection용 benchmarks의 발전
5. low-shot object detection
6. detection task에 적합한 backbone architecture
7. 기타의 다른 연구 주제
    1. large batch learning 또는 incremental learning

Before Reading

- 논문 링크: https://arxiv.org/abs/1908.03673

- 딥러닝을 이용한 visual object detection의 최근 발전(~2019)에 대한 포괄적인 내용을 담고 있음.
  - 필요에 따라, 딥러닝 이전의 연구들도 포함시킴.

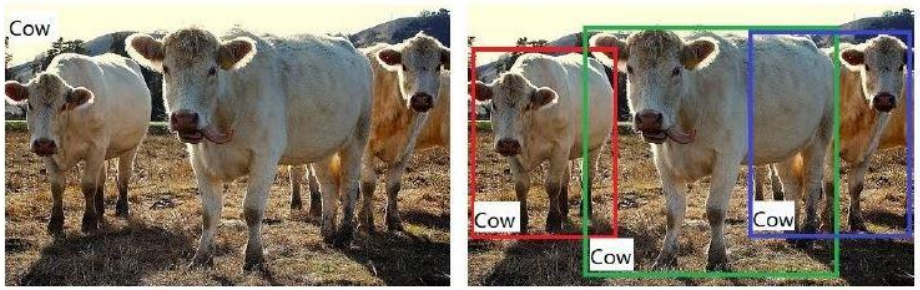
- 존재하는 object detection 프레임워크를 시스템적으로 분석하여 아래와 같이 크게 3가지 부분으로 나눔.
  - Detection Components
  - Learning Strategies
  - Applications & Detection Benchmarks

- 기타
  - 다양한 종류의 two-stage 및 one-stage detector들에 대한 상세한 설명은 하지 않음.
    - 기타 다양한 알고리즘들에 대해서도 중요한 부분만을 설명하였으며, 관심이 있을 경우 해당 논문의 링크를 참고하길 바람.

  - 알고리즘의 제안자가 알고리즘 이름을 약어로 사용하는 경우에만 표기함.
    - 각 알고리즘에 대한 논문 링크는 설명 시 바로 옆에 표시하여 상세한 내용을 바로 읽을 수 있도록 함.

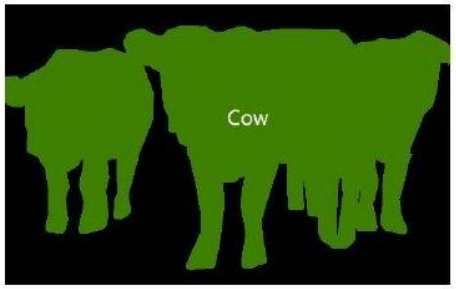
  - 최대한 다양한 분야의 사람들이 읽을 수 있도록 수식은 사용하지 않음.
    - 부담 없이 반복적으로 편하게 읽을 수 있으며, 본 글의 내용만 숙지하더라도 object detection task와 관련된 전반적인 지식 습득에 큰 도움이 될 것이라고 생각됨.

Comparison of different visual recognition tasks in computer vision

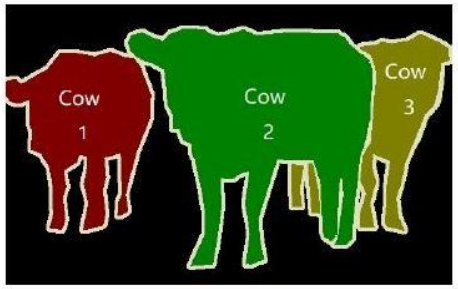


(a) Image Classification

(b) Object Detection



(c) Semantic Segmentation

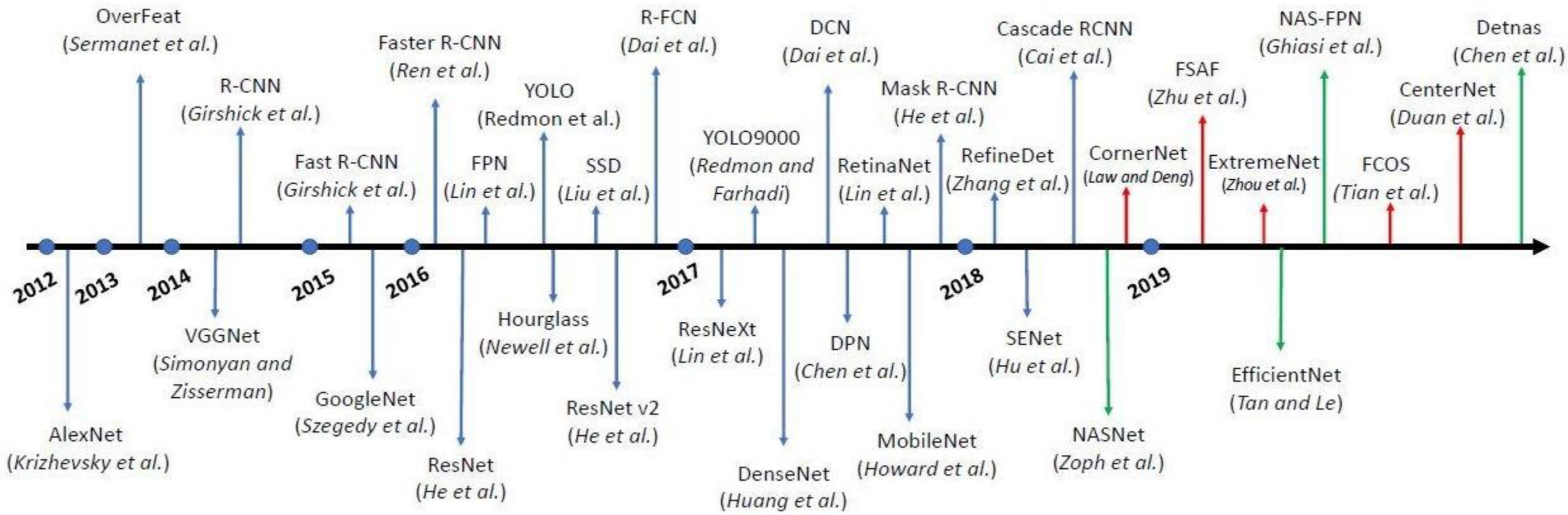


(d) Instance Segmentation

- (a) image classification
  - 이미지에 카테고리에 해당하는 class의 label을 할당하는 것을 목적으로 하는 task
- (b) object detection
  - 카테고리에 해당하는 label을 예측하는 것뿐만 아니라, bounding boxes를 이용하여 각각의 object instance에 대한 위치를 찾아야 하는(localization) task
- (c) semantic segmentation
  - object instance를 구별하지 않고, 각 픽셀별로 카테고리에 해당하는 label을 예측하는 것을 목적으로 하는 task
- (d) instance segmentation
  - object detection의 특별한 설정(setting)에 해당하며, 픽셀 수준(level)의 segmentation mask를 이용하여 다른 object instance와 구별을 하는 task

Major Milestone in Object Detection

- **빨강: Anchor-Free, 녹색: AutoML**
  - 이 둘은 향 후 연구의 잠재적인 주요 방향성을 제시



Taxonomy of key Methodologies

<table>
<tr><th colspan="4">Object Detection</th></tr>
<tr><th colspan="3">Detection Components</th><th>Learning Strategy</th><th>Applications & Benchmarks</th></tr>
<tr><th>Detection Settings</th><th>Detection Paradigms</th><th>Backbone Architecture</th><th>Training Stage</th><th>Applications</th></tr>
<tr><td rowspan="2">Bounding Box</td><td rowspan="2">Two-Stage Detectors</td><td rowspan="2">VGG16,ResNet,DenseNet</td><td>Data Augmentation</td><td rowspan="2">Face Detection</td></tr>
<tr><td>Imbalance Sampling</td></tr>
<tr><td rowspan="3">Pixel Mask</td><td rowspan="3">One-Stage Detectors</td><td rowspan="2">MobileNet, ResNeXt</td><td>Localization Refinement</td><td rowspan="2">Pedestrian Detection</td></tr>
<tr><td>Cascade Learning</td></tr>
<tr><td>DetNet, Hourglass Net</td><td>Others</td><td rowspan="2">Others</td></tr>
<tr><th colspan="2">Proposal Generation</th><th>Feature Representation</th><th>Testing Stage</th><th>Public Benchmarks</th></tr>
<tr><td colspan="2">Traditional Computer Vision Methods</td><td>Multi-scale Feature Learning</td><td rowspan="2">Duplicate Removal</td><td rowspan="2">MSCOCO, Pascal VOC, Open Images</td></tr>
<tr><td colspan="2">Anchor-based Methods</td><td>Region Feature Encoding</td></tr>
<tr><td colspan="2">Keypoint-based Methods</td><td>Contextual Reasoning</td><td rowspan="2">Model Acceleration</td><td>FDDB, WIDER FACE</td></tr>
<tr><td colspan="2">Other Methods</td><td>Deformable Feature Learning</td><td rowspan="2">Others</td><td rowspan="2">KITTI, ETH, CityPersons</td></tr>
</table>

Detection Components

1. Detection Settings & 2. Detection Paradigms

- **Detection Settings**
  - Bounding Box
    - vanilla object detection(bounding box 수준의 localization)에서 사용함.
    - 사각형 bounding boxes로 object를 localization하는 것을 목표로 하며, bounding box에 대한 annotation 정보만을 필요로 함.
    - 성능 평가 시 예측된 bounding box와 ground truth 사이의 IoU를 계산함.
  - Pixel Mask
    - instance segmentation(픽셀 수준 또는 mask 수준의 localization)에서 사용함.
    - 픽셀 별 mask로 각 object를 segment하는 것을 목표로 하며, 더욱 더 정밀한 픽셀 수준의 예측을 필요로 함.
    - 성능 평가 시 예측된 mask와 ground truth 사이의 IoU를 계산함.

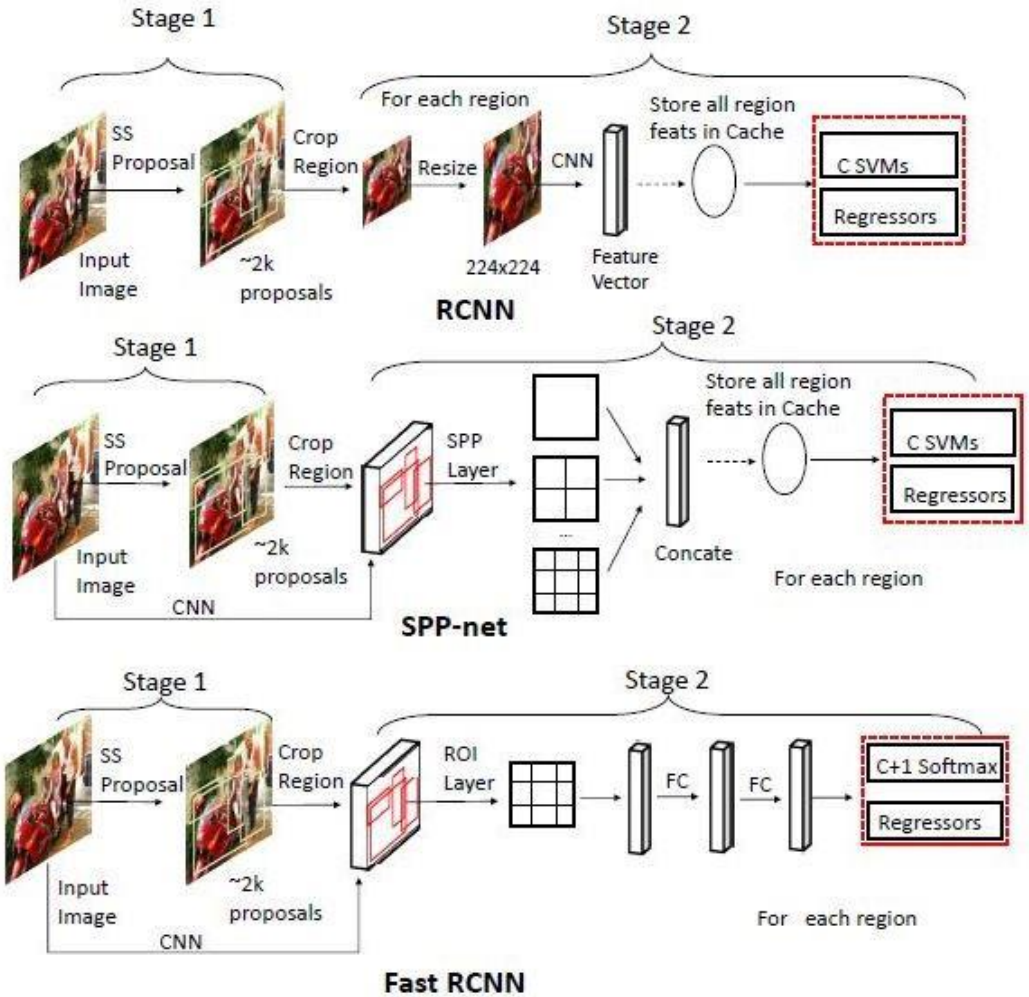
- **Detection Paradigms**
  - 현재 딥러닝 기반의 최신 object detector들은 크게 Two-Stage Detector와 One-Stage Detector들로 나뉨.
  - 다양한 종류의 Two-Stage 및 One-Stage Detectors들에 대한 상세한 알고리즘 설명은 해당 논문들을 참고하길 바람.

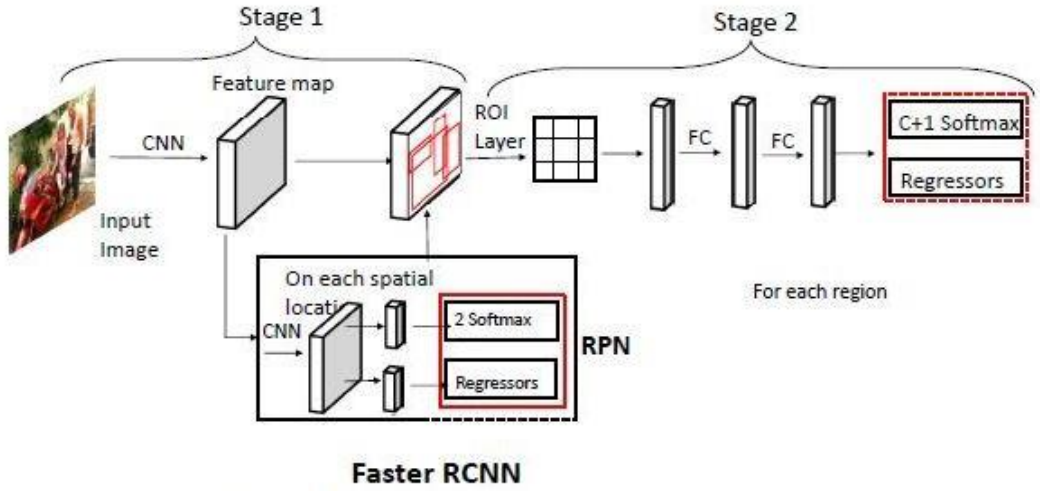
  - **1) Two-Stage Detectors**
    - 첫번째 단계에서 sparse한 proposal set을 생성하고, 두번째 단계에서는 생성된 proposal에 대한 feature vector를 deep convolutional neural networks를 이용하여 부호화시킨 후 object에 대한 class를 예측함.
    - 많은 공용 benchmark datasets에서 SOTA의 결과를 보이지만, 일반적으로 추론 속도가 느리다는 단점이 있음.
    - **출력: Bounding Box**
      - R-CNN(https://arxiv.org/abs/1311.2524)
      - SPP-net(https://arxiv.org/abs/1406.4729)
      - Fast R-CNN(https://arxiv.org/abs/1504.08083)
      - Faster R-CNN(https://arxiv.org/pdf/1506.01497.pdf)
      - R-FCN(https://arxiv.org/abs/1605.06409)
      - FPN(https://arxiv.org/abs/1612.03144)

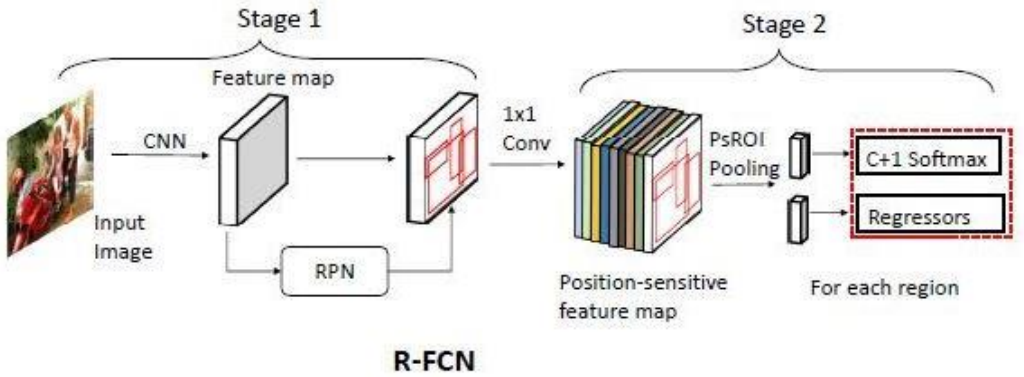
    - **출력: Pixel Mask(Instance Segmentation)**
      - Mask R-CNN(https://arxiv.org/abs/1703.06870)
      - Mask Scoring R-CNN(https://arxiv.org/abs/1903.00241)

- **Detection Paradigms**
  - **1) Two-Stage Detectors**

빨간색 점선 사각형: loss function이 정의된 출력



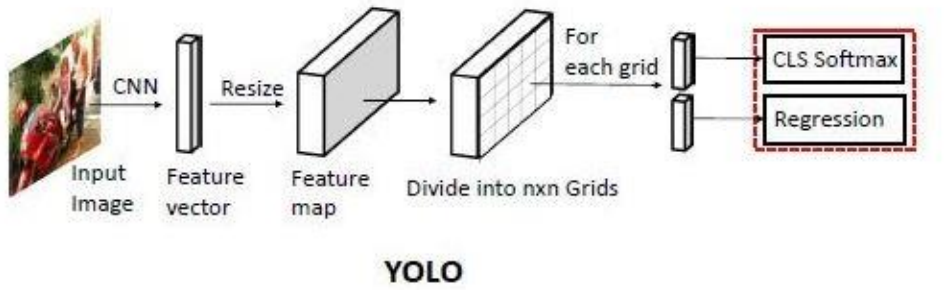




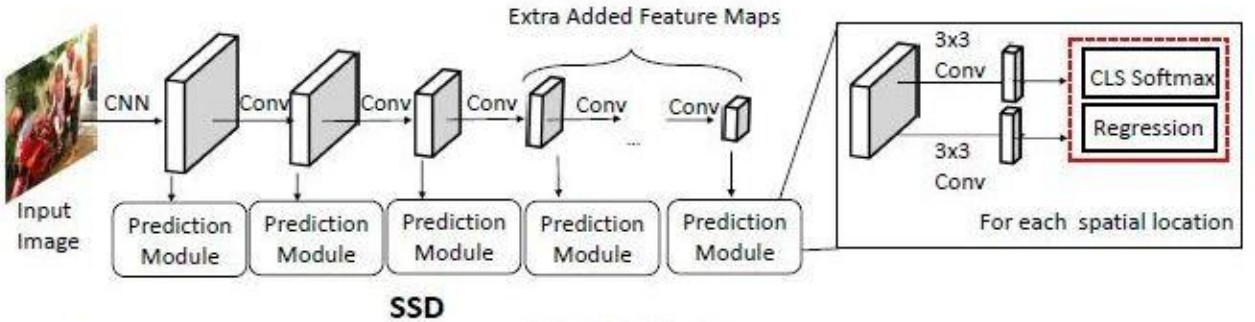
- **Detection Paradigms**
  - **2) One-Stage Detectors**
    - proposal을 생성하는 단계(proposal 생성을 위한 학습을 의미)를 따로 분리하지 않음.
    - 이미지 내 모든 위치를 잠재적인 후보 object들로 고려하여 각각의 관심 영역을 배경 또는 타겟 object들로 분류하도록 함.
    - two-stage detector들에 비해서 훨씬 더 빠르므로 실시간 object detection에 적합하지만, 상대적으로 저조한 성능을 보임.
    - **출력: Bounding Box**
      - OverFeat(https://arxiv.org/abs/1312.6229)
      - YOLO(https://arxiv.org/abs/1506.02640)
      - SSD(https://arxiv.org/abs/1512.02325)
      - RetinaNet(https://arxiv.org/abs/1708.02002)
      - YOLOv2(https://arxiv.org/abs/1612.08242)
      - CornerNet(https://arxiv.org/abs/1808.01244)

- **Detection Paradigms**
  - **2) One-Stage Detectors**

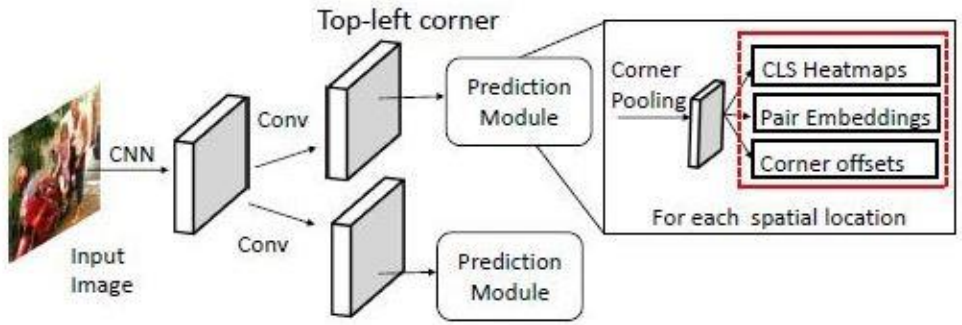
빨간색 점선 사각형: loss function이 정의된 출력



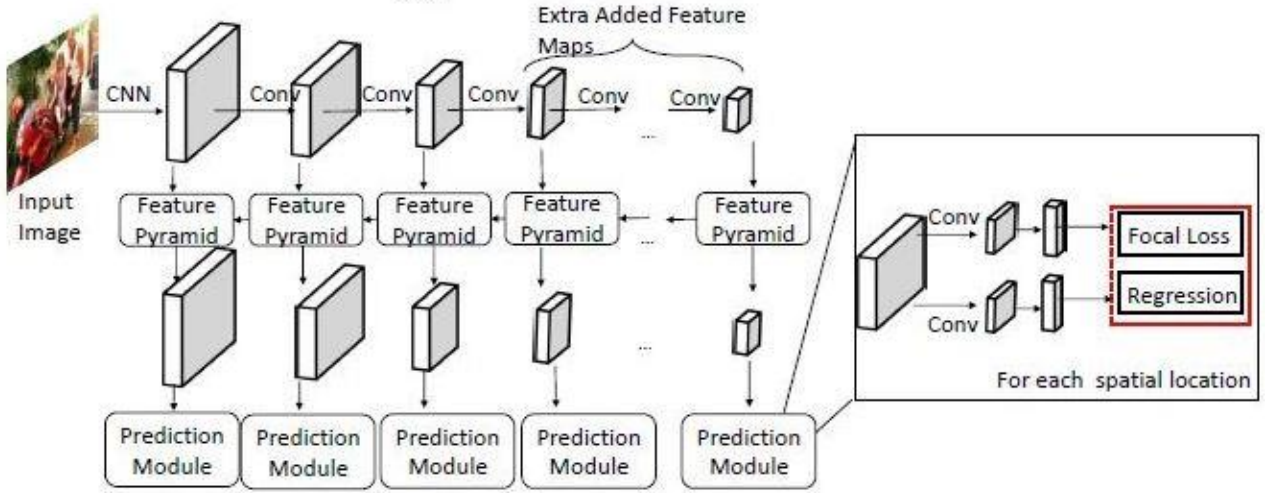
YOLO



SSD



CornerNet



RetinaNet

3. Backbone Architecture

- **Basic Architecture of a CNN**
  - 연속적인 Convolution Layers → Non-Linear Activation Layers → Pooling Layers → Fully Connected Layers
  - AlexNet(https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf): 전형적인 convolutional neural network

- **CNN Backbone for Object Detection**
  - **image classification과 object detection을 위해 고안된 네트워크의 차이**
    - VGG16, ResNet, ResNet-v2, DenseNet, DPN, ResNeXt, MobileNet, GoogleNet 등의 네트워크는 image classification을 위해 고안됨.
    - DetNet는 object detection을 위해서, Hourglass Network는 human pose estimation을 위해서 고안됨.
    - image classification을 위해 고안된 네트워크를 object detection에 이용하기 위해서는 전형적으로 ImageNet을 이용하여 훈련시킴.
      - 이렇게 classification으로부터 pre-trained 모델을 직접적으로 적용하는 것은 classification과 detection 사이의 잠재적인 충돌(conflict)로 인해 차선책(sub-optimal)적인 방법임.
      - i) classification은 공간 불변성(spatial invariance)을 유지하기 위해 큰 receptive fields를 필요로 하므로, feature map의 해상도를 낮추기 위해 pooling layer와 같은 여러 번의 downsampling 연산을 적용함.
        - 따라서, 생성된 feature map은 저해상도의 공간적으로 불변하는 큰 receptive fields를 갖게 됨.
        - 하지만, detection은 object의 위치를 올바르게 찾기 위해 고해상도의 공간 정보를 필요로 함.
      - ii) classification은 하나의 feature map에서 예측을 수행하지만, detection은 다중 스케일에서 object를 검출해야 하므로 multiple representations이 가능한 feature map을 필요로 함.

  - **고안 목적: image classification**
    - VGG16(https://arxiv.org/abs/1409.1556)
      - AlexNet을 기반으로 한 방법임.
      - convolutional layers를 쌓아 네트워크의 깊이를 증가시킴으로써, 모델의 expression capability가 증가되고, 더 나은 성능을 보이는 것을 확인함.
      - convolutional layers 단순히 쌓아서 layers의 깊이를 20까지 증가시킬 경우, SGD를 이용한 최적화 문제에 어려움을 보임.

    - ResNet(https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf)
      - VGG16의 최적화 문제를 shortcut connections을 도입하여 해결하고자 함.
      - shortcut connections은 deep layers로부터 shallow units까지 gradient를 직접적으로 전파할 수 있는 highway를 만들어주므로 훈련의 어려움을 상당히 줄일 수 있음.
      - residual blocks을 이용하여 네트워크를 효율적으로 훈련시킴으로써, 모델의 깊이는 16에서 152까지 증가하였고, 상당히 높은 수용력을 가지는 모델을 만들 수 있음.

- **CNN Backbone for Object Detection**
  - **고안 목적: image classification**
    - ResNet-v2(https://arxiv.org/abs/1603.05027)
      - Batch Normalization를 적절하게 배치하므로써, 기존 ResNet보다 뛰어난 성능을 보임.
      - 1000개 이상의 layers를 성공적으로 학습시킬 수 있었으며, 늘어난 네트워크의 깊이를 통해 여전한 성능 향상을 시킬 수 있었음.

    - DenseNet(https://arxiv.org/abs/1608.06993)
      - ResNet이 shortcut connections을 적용하여 훈련의 어려움을 줄일 수 있었지만, 이전 layer로부터 온 특징들을 완전히 활용하지 못한다는 점에 착안함.
      - ResNet의 shallow layers 내 원래 특징은 element-wise 연산 시 잃어버리게 되므로, 추 후에 직접적으로 사용할 수 없음.
      - ResNet의 element-wise 덧셈을 사용하는 대신에 입력과 residual 출력을 연결시킴으로써(concatenating) shallow layers 내 특징을 유지하면서, 정보 흐름을 개선시킴.

    - DPN(https://arxiv.org/abs/1707.01629)
      - DenseNet에서 shallow layers로부터 새롭게 얻어낸 특징들의 상당수는 중복적이므로, 상당한 계산 비용이 발생함.
      - ResNet과 DenseNet의 장점을 통합하는 방식이 아니라, 입력 채널을 2개의 부분으로 나누어 계산 후 최종 출력을 연결시키는 방식임.

    - ResNeXt(https://arxiv.org/abs/1611.05431)
      - 비교가능한 분류 정확도를 유지하면서 계산 및 메모리 비용을 상당히 줄일 수 있는 방법임.
      - 계산 비용을 줄이기 위해 feature map channels을 sparse하게 연결하는 group convolution layers를 적용함.
      - 원래 ResNet과 계산 비용을 일관성 있게 유지하기 위해 group의 개수를 증가시킴으로써, 훈련 데이터로부터 semantic feature representation을 풍부하게 capture하고 backbone의 정확도를 향상시킴.

    - MobileNet(https://arxiv.org/abs/1704.04861)
      - 각 feature map의 channel 개수와 동일하게 좌표(coordinates)를 설정한 방법으로서, 모바일 플랫폼을 위해 고안된 방법임.
      - 분류 정확도의 상당한 손실 없이도, 계산 비용 및 파라미터의 개수를 상당히 줄일 수 있는 방법임.

- **CNN Backbone for Object Detection**
  - **고안 목적: image classification**
    - GoogleNet
      - 초기 버전(Inception-v1, https://arxiv.org/abs/1409.4842)
        - 모델의 깊이를 증가시키는 것 외에도, 학습 수용력을 향상시키기 위해 모델의 너비를 늘리는 이점과 관련된 연구임.
        - 주어진 layer 내 동일한 feature map에 대해 서로 다른 스케일의 convolution kernels를 적용하는 inception module을 제안함.
        - 이를 통해 다중 스케일의 features를 capture할 수 있으며, 이러한 feature들은 출력 feature map과 함께 요약할 수 있음.
      - 변종 버전(Inception-v2~v4)
        - convolution kernels의 선택을 위한 설계 방법에 대한 연구(Inception-v2~v3, https://arxiv.org/abs/1512.00567)
        - residual blocks을 도입한 연구(Inception-v4, https://arxiv.org/abs/1602.07261)

  - **고안 목적: object detection**
    - DetNet(https://arxiv.org/abs/1804.06215)
      - classification task과 detection task 사이의 연결을 위해 제안된 방법으로서 특히, detection을 위해서 고안된 방법임.
      - 예측을 위한 feature map을 고해상도로 유지하기 위해, dilated convolutions을 이용하여 receptive fields를 증가시킴.
      - 풍부한 정보를 제공하는 다중 스케일 상에서 object를 검출하며, 대규모의 classification 데이터셋을 이용하여 pretrained 되었고 네트워크의 구조는 검출을 위해 설계되어 있음.

  - **고안 목적: human pose recognition**
    - Hourglass Network(https://arxiv.org/abs/1603.06937)
      - 사람의 자세 인식을 위해 처음 등장한 네트워크로서, 연속적인 hourglass modules을 가진 fully convolutional 구조로 되어 있음.
      - 연속적인 convolutional layer 또는 pooling layer를 통해 입력 이미지를 downsampling 한 후, deconvolutional 연산을 통해 feature map을 upsampling 함.
      - downsampling 시 정보 손실을 피하기 위해, downsampling features와 upsampling features 간에 skip connection을 적용함.
      - 지역 및 전역 정보를 모두 capture 할 수 있기 때문에 object detection에 상당히 적합함.
      - 현재 SOTA(State-Of-The-Art) detection 프레임워크에서 가장 널리 사용되고 있음.

4. Proposal Generation

- 잠재적으로 object가 될 수 있는 사각형 bounding boxes의 set을 생성하며, object detection 프레임워크에서 상당히 중요한 역할을 함.

- classification과 localization의 정제(refinement)를 위해서 사용됨.

- 전통적인 컴퓨터 비전 기법들, anchor 기반의 지도 학습(supervised learning) 기법들, keypoint 기반 기법들, 기타 기법들로 분류함.

- one-stage detector 및 two-stage detector 모두 proposal을 생성하나 아래와 같은 차이가 있음.
  - two-stage detector의 경우 전경 또는 배경의 정보를 가지는 sparse한 proposal set을 생성함.
  - one-stage detector의 경우 이미지 내 각 region을 잠재적인 proposal로 여기고, 각 위치에서의 잠재적인 object에 대한 class와 bounding box 좌표 정보를 함께 추정함.

- 사견: "What makes for effective detection proposals?, https://arxiv.org/pdf/1502.05082.pdf)"은 다양한 종류의 detection proposals에 대한 성능 평가를 한 논문으로 자세한 내용이 필요할 경우 리뷰하는 것을 권장함.

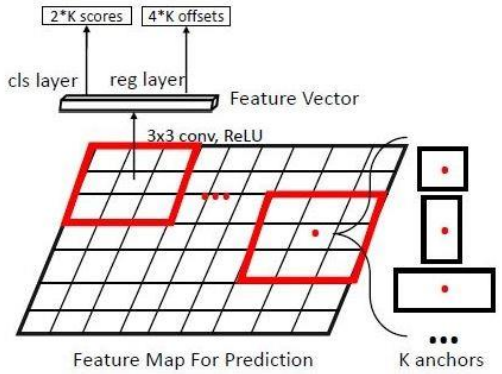
- **1) 전통적인 컴퓨터 비전 기법들**
  - edge, corner, color 등과 같은 low-level의 cue를 이용하여 proposal을 생성함.
  - 단순하며, Pascal VOC와 같은 중간 크기의 dataset에서는 높은 recall을 보이는 proposal을 생성할 수 있다는 이점이 있음.
  - 하지만, color나 edge와 같은 low-level의 시각적 cue에 기반하고 있기 때문에 전체 detection 파이프라인과 함께 최적화 될 수 없음.
  - 표현 학습(representation learning)의 개선을 위한 대규모 dataset의 능력을 활용할 수 없음.
  - 이러한 한계들로 인해 MSCOCO와 같은 도전적인 dataset에 대해서 높은 수준의 proposal을 생성하는데 어려움을 겪게 됨.

  - 후보 box에 대해 objectness score를 계산하는 방법, 원래 이미지로부터 super-pixel을 merge하는 방법, 여러개의 전경 및 배경 분할을 생성하는 방법 등으로 나뉨.
  - **i) 후보 box에 대해 objectness score를 계산하는 방법**
    - 각각의 후보 box가 object가 될 수 있는 가능성을 score로 측정하여 예측함.
    - 관련 연구 1(https://groups.inf.ed.ac.uk/calvin/Publications/alexe12pami.pdf)
      - color contrast, edge density, saliency와 같은 visual cue들에 기반한 분류를 함으로써, proposal에 대해 objectness score를 할당함.
    - 관련 연구 2(http://www.ee.oulu.fi/~jkannala/publications/iccv2011.pdf)
      - 관련 연구 1을 기반으로 하고 있으며, 후보 proposal의 objectness score의 순위를 매기기 위해 보다 효율적인 cascaded learning을 도입함.

  - **ii) 원래 이미지로부터 super-pixel을 merge하는 방법**
    - 분할(segmentation) 결과로부터 생성된 superpixel들을 merge하는 것을 기반으로 함.
    - 관련 연구 1(Selective Search, http://www.huppelen.nl/publications/selectiveSearchDraft.pdf)
      - super-pixels의 merge에 기반한 proposal 생성 알고리즘임.
      - 분할 기법(http://people.cs.uchicago.edu/~pff/papers/seg-ijcv.pdf)을 이용하여 생성된 여러개의 계층적인 segments를 계산하고, 이들의 color, areas 등과 같은 시각적인 factor에 따라서 merge를 함.
      - merge된 결과에 최종적인 bounding boxes가 배치됨.
    - 관련 연구 2(https://www.vision.ee.ethz.ch/publications/papers/proceedings/eth_biwi_01061.pdf)
      - Selective Search와 유사한 아이디어를 사용함.
      - 차이점으로는 merging function에 대한 weight를 학습시켰고, merge가 random하게 처리가 됨.
    - Selective Search는 다른 전통적인 방법들에 비해서 높은 recall과 효율성 때문에 널리 사용되고 있음.

- **1) 전통적인 컴퓨터 비전 기법들**
  - **iii) 여러개의 전경 및 배경 분할을 생성하는 방법**
    - 여러개의 seed region으로부터 출발하여, 각 seed에 대해 전경 및 배경에 해당하는 segment를 생성함.
    - 관련 연구 1(CPMC, https://www.researchgate.net/publication/51855547_CPMC_Automatic_Object_Segmentation_Using_Constrained_Parametric_Min-Cuts)
      - 계층적 분할을 피하기 위해 다양한 seed를 이용하여 초기화된 겹쳐진 segment set을 생성함.
      - 각각의 proposal segment는 전경 또는 배경에 해당하는 이진 분할 문제를 해결하는 해(solution)가 됨.
    - 관련 연구 2(http://dhoiem.cs.illinois.edu/publications/pami2013_proposals_endres.pdf)
      - Selective Search와 CPMC의 아이디어를 결합한 방식임.
      - super-pixels로 시작하여 이들을 새롭게 고안된 features를 이용하여 merge함.
      - 이렇게 merge된 segment는 보다 큰 segment를 생성하기 위한 seed로 사용되며, 이는 CPMC와 유사한 방법임.
      - 하지만, 고품질의 segmentation mask를 생성하는 것은 시간이 많이 소요되며, 대규모의 dataset에는 적용이 어려움.

- **2) anchor 기반 지도 학습 기법들**
  - 지도 학습을 이용한 proposal 생성 방법의 큰 부류는 anchor를 기반으로 하는 접근들이며, 미리 정의된 anchor를 기반으로 proposal을 생성함.
    - **고정된 anchor를 이용하는 방법들**
      - RPN(https://arxiv.org/pdf/1506.01497.pdf)
        - deep convolutional feature maps을 기반으로 지도 학습을 이용하여 proposal을 생성하는 방법임.
        - 3 x 3 크기의 convolution filters를 이용하여 전체 feature map 상에서 sliding 시킴.
        - 각 위치에 대해 다양한 크기 및 종횡비를 가진 k개의 anchor(또는 bounding boxes의 초기 추정치)가 고려되며, 전체 이미지 상에서 서로 다른 스케일을 가진 object들을 매칭하는데 사용됨.
        - ground truth bounding boxes에 기반하여, object의 위치는 anchor 추정을 위한 지도 신호(supervision signal)을 얻도록 가장 적합한 anchor에 매칭이 됨.
        - 각 anchor로부터 추출된 256 차원의 특징 벡터는 2개의 branch로 나뉘게 됨.
          - classification branch: objectness score를 모델링하는 역할을 담당함.
          - regression branch: 원래 anchor 추정치로부터 bounding box의 위치를 정제하기 위해 4개의 실수값을 부호화시킴(encoding).
        - ground truth에 기반하여, classification branch에 따라 각 anchor는 object 또는 배경으로 예측이 되어짐.
        - feature map의 각 위치는 sliding window와 연결되어 있고, 2개의 sibling branch가 이어져 있음(아래 그림 참고).



      - SSD(https://arxiv.org/abs/1512.02325)
        - object를 매칭하기 위해 다중 스케일의 anchor를 이용하는 RPN과 유사한 방법으로 anchor를 이용함.
        - RPN과의 차이점
          - RPN은 먼저 anchor proposal이 배경 또는 전경인지를 평가한 후 다음 단계에서 카테고리에 해당하는 classification를 수행함.
          - SSD는 각 anchor proposal에 카테고리에 해당하는 확률을 할당함.

- ## 2) anchor 기반 지도 학습 기법들
  - **anchor 설계를 개선한 방법들**
    - 우수한 성능에도 불구하고 anchor priors는 여전히 다중 스케일과 종횡비를 이용하여 휴리스틱한 방법으로 수동적으로 설계가 되어 있음.
    - 이러한 설계가 최적의 선택은 아니며, 다른 dataset은 다른 anchor 설계 전략을 필요로 하기 때문에 이를 개선하기 위한 다양한 연구들이 진행됨.
    - 관련 연구 1(S3FD, https://arxiv.org/abs/1708.05237)
      - object를 매칭하기 위해 주의깊게 설계된 anchor를 이용하는 SSD를 기반으로 함.
      - 서로 다른 feature map의 효율적인 receptive field에 따라서, 서로 다른 anchor prior가 설계됨.
    - 관련 연구 2(https://arxiv.org/abs/1802.09058)
      - 입력 이미지의 크기를 크게 하고 anchor stride를 줄임으로써 작은 object를 매칭할 수 있도록 설계된 anchor를 도입함.
    - 관련 연구 3(DeRPN, https://arxiv.org/abs/1811.06700)
      - RPN에 기반하여 anchor boxes의 차원을 분해함.
      - anchor string 메커니즘을 적용하여 object의 높이 및 너비를 독립적으로 매칭시킴.
      - 이러한 메커니즘은 대규모의 분산을 가진 object를 매칭시키고 탐색 공간을 줄이는데 도움이 됨.
    - 관련 연구 4(DeepProposals, https://arxiv.org/abs/1606.04702)
      - 저해상도의 deeper layer를 가진 feature map 상에서 proposal을 예측함.
      - 그 다음, 고해상도의 shallow layer를 가진 feature map으로 역투영 되고 정제가 됨.
    - 관련 연구 5(YOLOv2, https://arxiv.org/abs/1612.08242)
      - k-means clustering을 이용하여 훈련 데이터로부터 priors를 학습시키는 방법으로 anchor priors를 설계함.
    - 관련 연구 6(RefineDet, https://arxiv.org/abs/1711.06897)
      - 수동으로 정의된 anchor를 2단계로 정제함.
      - 첫번째 단계에서는 원래 수동으로 고안한 anchor를 기반으로 localization offset의 set을 학습시킨 후 학습된 offset을 이용하여 anchor를 정제시킴.
      - 두번째 단계에서는 향 후 정제를 위해, 첫번째 단계에서 정제된 anchor를 기반으로 새로운 localization offset의 set을 학습시킴.
      - 이렇게 cascaded된 최적화 프레임워크를 이용하여 data-driven 방법으로 anchor의 품질과 최종 예측 정확도를 상당히 개선시킴.
    - 관련 연구 7(Cascade R-CNN, https://arxiv.org/abs/1712.00726)
      - RefineDet과 유사하게 cascaded 방식으로 proposal을 정제하는 방식을 적용함.
    - 관련 연구 8(MetaAnchor, https://arxiv.org/abs/1807.00980)
      - customized된 anchors로부터 계산된 신경망에 의해 구현된 function으로 anchor를 모델링함.
      - 수동적으로 정의된 anchor를 이용하는 다른 방법들 보다 포괄적인 성능 개선을 보였지만, customized된 anchors는 여전히 수동으로 설계가 되어 있음.

3) keypoint 기반 기법들(anchor-free 기법들)

- anchor-free 기법들은 유망한 향 후 연구 방향 중 하나임.

- keypoint를 기반으로 proposal을 생성하는 방법으로서, corner 기반 기법들과 center 기반 기법들의 두가지 부류로 나뉨.
- **i) corner 기반 기법들**
  - feature map으로부터 학습된 corner 쌍을 merge함으로써 bounding boxes를 예측하는 방법임.
  - 관련 연구 1(DeNet, https://arxiv.org/abs/1703.10295)
    - object detection 문제를 확률적인 방법으로 개조함.
    - feature map 상의 각 point에 대해, object의 4개의 corner type(좌상단, 우상단, 좌하단, 우하단) 중에 하나가 되는 분포를 모델링함.
    - 그리고, object의 각 corner에 대해 naive bayesian classifiers를 적용하여 bounding box의 confidence score를 추정함.
    - anchor 설계를 제거하기 때문에 더욱 더 효율적으로 고품질의 proposal을 생성할 수 있는 방법임.
  - 관련 연구 2(CornerNet, https://arxiv.org/abs/1808.01244)
    - DeNet에 기반한 방법으로서, corner 상에서 카테고리에 해당하는 정보를 직접적으로 모델링함.
    - 새로운 feature embedding 방법과 corner pooling layer를 이용하여 좌상단 및 우하단 corner에 대한 정보를 모델링함.
    - 이를 통해 동일한 object에 속하는 keypoint들을 정확하게 매칭할 수 있으며, 공용 벤치마크에서 SOTA의 결과를 얻음.

- **ii) center 기반 기법들**
  - feature map 상의 각 위치에 대해서 object의 center가 될 확률을 예측하며, 높이 및 너비는 어떠한 anchor priors도 없이 직접적으로 regression됨.
  - 관련 연구 1(FSAF, https://arxiv.org/abs/1903.00621)
    - FPN 구조를 이용하는 one-stage detector에 접목할 수 있는 프레임워크로서, feature pyramid의 각 level에 부착된 multi level의 center 기반 branch들을 훈련시킬 수 있도록 온라인의 feature selection block이 적용됨.
    - 훈련 동안에, center 기반 branch를 훈련시키기 위해 각 object에 가장 적합한 feature level을 동적으로 할당함.
  - 관련 연구 2(https://arxiv.org/abs/1904.07850)
    - FPN 구조가 없는 단일 Hourglass network을 이용하여 제안된 새로운 center 기반 프레임워크임.
    - 3D detection, 사람 자세 인식 등과 같은 고수준의 문제에 center 기반 기법을 적용하였으며, SOTA의 결과를 보임.
  - 관련 연구 3(CenterNet, https://arxiv.org/abs/1904.08189)
    - center 기반 기법과 corner 기반 기법을 결합한 방법으로써, baseline 방법들보다 상당한 개선을 이룸.
    - corner 쌍을 이용하여 bounding boxes를 예측한 후, 쉬운 negative를 제거하기 위해 초기 예측치가 center가 될 확률을 예측함.

- **4) 기타 기법들**
  - keypoint 또는 anchor에 기반하지 않으면서 proposal을 생성하는 접근법들이 있으며, 경쟁력 있는 결과를 보임.
  - 관련 연구 1(AZnet, https://arxiv.org/abs/1512.07711)
    - 관심이 높은 영역(regions of high interest)에 자동적으로 초점을 맞추는 방법임.
    - object가 있을 가능성이 있는 sub-region에 적응적으로 계산 자원을 향하게 하는 탐색 전략을 적용함.
    - 각 region에 대해 2개의 값을 예측함.
      - zoom indicator: 더 작은 object를 포함하도록 region을 더욱 더 나눌지 여부를 결정함.
      - adjacency scores: zoom indicator의 objectness를 나타냄.
    - 전체 이미지를 시작점으로 하여, 각각의 분할된 sub-region은 zoom indicator가 너무 작아질 때가지 위의 방법을 이용하여 재귀적으로 처리가 됨.
    - sparse하며 작은 object를 매칭하는데 있어서, RPN의 anchor와 object를 매칭하는 접근 보다 효과적임.

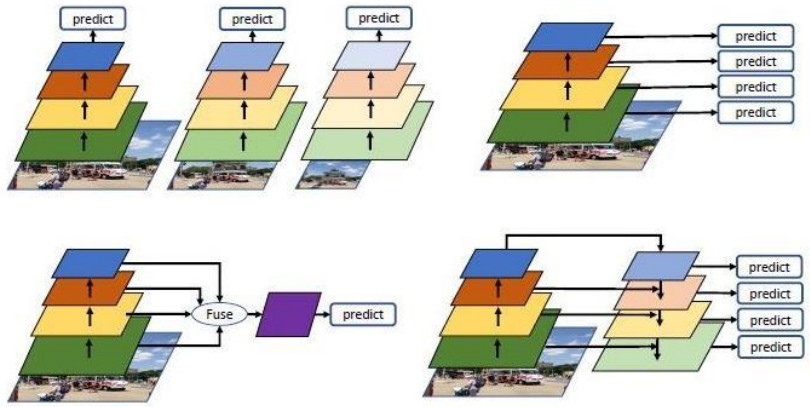
5. Feature Representation Learning

- 특징 표현 학습은 전체 detection 프레임워크에서 중요한 구성요소로서, 일반적으로 검출 대상 object는 복잡한 환경에서 스케일과 종횡비의 차이가 큰 경우가 많음.

- 좋은 검출 성능을 보이기 위해서는 object를 잘 표현하도록 강건하며, 변별력 있는 특징을 학습시켜야 함.

- 특징 표현 학습을 위한 전략은 크게 다중-스케일 특징 학습(multi-scale feature learning), 영역 특징 부호화(region feature encoding), 문맥 추론(contextual reasoning), 변형 가능한 특징 학습(deformable feature learning) 등으로 나뉨.

- ## 1) 다중-스케일 특징 학습(multi-scale feature learning)
  - Fast R-CNN, Faster RCNN 등과 같이 deep convolutional networks에 기반한 전형적인 object detection은 object를 검출하기 위해 single layer의 feature map만을 사용함.
  - single feature map 상에서 다양한 범위의 스케일과 종횡비를 갖는 object를 검출하는 것은 상당히 어려운 일임.
  - deep convolutional networks는 서로 다른 스케일 정보를 capture하기 위해 서로 다른 layer에서 계층적으로 특징을 학습시킴.
    - shallow layer features: 공간적으로 풍부한 정보를 가지고 있으며, 더 높은 해상도와 더 작은 receptive fields를 가지고 있으므로 작은 object들을 검출하는데 좀 더 적합함.
    - deep layer features: 의미적으로 풍부한 정보를 가지고 있으며, 조명, 이동 등에 보다 강건하고, 더 작은 해상도와 더 큰 receptive fields를 가지고 있으므로 큰 object들을 검출하는데 좀 더 적합함.
    - 작은 object들을 검출 할 경우, 고해상도의 표현이 필요하며 작은 object들에 대한 표현은 deep layer feature에서는 사용할 수 없으므로 작은 object들의 검출을 어렵게 만듬.
      - dilated/atrous convolutions(R-FCN(https://arxiv.org/abs/1605.06409), DCN(https://arxiv.org/abs/1703.06211)) 등의 기법이 다운샘플링을 피하기 위해 제안되었으며, 좀 더 깊은 layer에서 고해상도의 정보를 사용함.
    - shallow layer에서 큰 object들을 검출하는 것은 충분히 큰 receptive fields 없이는 최적으로 수행할 수 없음.
    - 따라서, 특징 스케일 문제를 다루는 것은 object detection 에서 근본적인 연구 주제가 되었음.
    - 다중 스케일 특징 학습 문제의 해결을 위한 4가지 주요 패러다임은 아래 그림처럼, 이미지 피라미드(Image Pyramid), 예측 피라미드(Prediction Pyramid), 통합 특징(Integrated Features), 특징 피라미드(Feature Pyramid) 등이 있음.



    - 좌상단: 이미지 피라미드 ➔ 서로 다른 스케일 이미지로부터 생성된 여러개의 detector를 학습시킴.
    - 우상단: 예측 피라미드 ➔ 여러개의 feature map 상에서 예측을 수행함.
    - 좌하단: 통합 특징 ➔ 여러개의 feature들로부터 생성된 하나의 feature map 상에서 예측을 수행함.
    - 우하단: 특징 피라미드 ➔ 예측 피라미드와 통합 특징을 결합한 구조임.

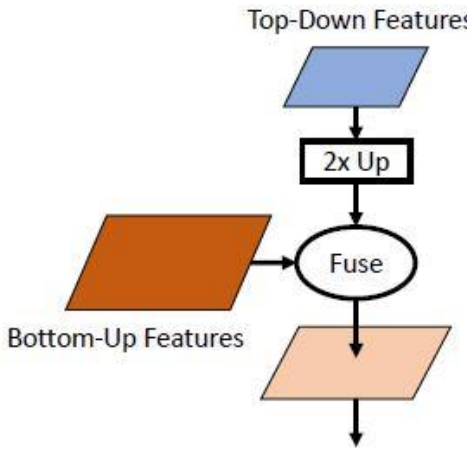
- ## 1) 다중-스케일 특징 학습(multi-scale feature learning)
  - **이미지 피라미드(Image Pyramid)**
    - 입력 이미지를 여러 개의 스케일을 갖도록 resize하고(이미지 피라미드), 여러개의 detector를 학습시키며, 각각의 detector들은 특정한 범위의 스케일에 대한 책임을 지고 있음.
    - 테스트 시, 여러개의 detector에 따라서 서로 다른 스케일로 이미지가 resize되고, detection 결과가 최종적으로 merge되며 이러한 방식은 계산량이 많음.
    - 관련 연구 1(https://arxiv.org/abs/1707.09531)
      - 모든 object를 유사한 스케일을 갖도록 이미지를 resize시켜 경량의 스케일을 알고 있는(light-weight scale aware) 네트워크를 학습시킴.
      - 그 다음에 단일 스케일의 detector를 학습시킴.
    - 관련 연구 2(SNIP, http://openaccess.thecvf.com/content_cvpr_2018/papers/Singh_An_Analysis_of_CVPR_2018_paper.pdf)
      - 작은 크기의 object를 detection하기 위한 포괄적인 실험을 수행함.
      - 모든 스케일의 object를 다루기 위해 단일 스케일의 강건한 detector를 학습시키는 것은 이미지 피라미드를 이용하여 스케일에 종족적인 detector를 학습시키는 것 보다 훨씬 더 어렵다고 주장함.
      - 여러개의 스케일에 종족적인 detector를 훈련시켰으며, 각 detector들은 특정한 스케일의 object에 대한 책임을 지게 됨.

  - **통합 특징(Integrated Features)**
    - 여러개의 layer에 대한 features들을 결합하여 하나의 feature map을 구성하고 해당 feature map에서 최종적인 예측을 수행하는 접근임.
    - 공간적으로 풍부한 shallow layer feature들과 의미적으로 풍부한 deep layer feature들을 융합함으로써, 새롭게 구성된 feature들은 풍부한 정보를 포함하고 있게 되며, 서로 다른 스케일의 object들을 검출할 수 있게 됨.
    - 이러한 조합은 주로 skip connections(https://arxiv.org/abs/1512.03385)을 이용하여 달성할 수 있음.
    - 서로 다른 layer의 feature norm은 높은 분산을 갖기 때문에 특징 정규화(feature normalization)가 필요함.
    - 관련 연구 1(ION, https://arxiv.org/abs/1512.04143)
      - ROI Pooling(https://arxiv.org/abs/1504.08083)을 통해 서로 다른 layer들에서 region feature를 crop하고, 이렇게 생성된 다중 스케일 region feature들을 결합하여 최종적인 예측을 수행함.
    - 관련 연구 2(HyperNet, https://arxiv.org/abs/1604.00600)
      - ION과 유사한 아이디어를 적용하였으며, proposal을 생성하고 object를 검출하기 위해 중간 및 shallow layer의 feature들을 통합함으로써 고해상도의 hyper feature map을 주의 깊게 고안함.
      - deep layer feature maps을 업샘플링 하기 위해 deconvolutional layers를 사용하였으며, 입력 blob을 정규화 하기 위해 batch normalization layers가 사용됨.
      - 구성된 hyper feature maps은 서로 다른 layer들로부터 생성된 문맥 정보(contextual information)를 함축적으로 부호화 함.
    - 관련 연구 3(MLKP, https://arxiv.org/abs/1804.00428)
      - object proposals을 위한 단순한 1차 표현을 이용하는 대신, 고차(high-order) 표현을 통합하는 세분화된 분류 알고리즘에 영감을 얻음.
      - proposal feature의 고차 통계(high-order statistics)를 capture하여, 더욱 더 변별력 있는 feature 표현을 효율적으로 생성함.
      - 결합된 feature 표현은 더욱 더 설명력을 가지며(descriptive) classification과 localization 모두를 위한 의미 및 공간 정보를 제공하게 됨.

- **1) 다중-스케일 특징 학습(multi-scale feature learning)**
  - **예측 피라미드(Prediction Pyramid)**
    - 관련 연구 1(SSD, https://arxiv.org/abs/1512.02325)
      - 여러개의 layer에서 생성된 coarse한 features와 fine한 features들을 함께 결합하는 방식임.
      - 여러개의 layer로부터 예측이 수행되며, 각각의 layer들은 특정한 scale의 object들에 대한 책임을 지게 됨.
      - 추후, 다중 스케일의 object 검출을 위해 다양한 연구들이 이러한 원리를 바탕으로 수행되어짐.
    - 관련 연구 2(http://www-personal.umich.edu/~wgchoi/SDP-CRC_camready.pdf)
      - 특정한 스케일에 해당하는 object proposal을 생성하기 위해 적절한 feature map을 활용하며, 이러한 feature map은 object를 예측하기 위해 여러개의 스케일에 종족적인(scale-dependent) classifiers로 전해짐.
      - 검출 속도를 빠르게 하기 위해, 초기 단계에서 쉬운 배경 proposal들을 제거할 수 있도록 cascaded rejection classifiers를 학습시킴.
    - 관련 연구 3(MSCNN, https://arxiv.org/abs/1607.07155)
      - 해상도 개선을 위해 deconvolutional layers를 여러개의 feature map에 적용하고, 이렇게 정제된 feature map들은 예측을 위해 추후에 사용됨.
    - 관련 연구 4(RFBNet, https://eccv2018.org/openaccess/content_ECCV_2018/papers/Songtao_Liu_Receptive_Field_Block_ECCV_2018_paper.pdf)
      - RFB(Receptive Field Block)를 통해 강건함과 receptive field를 향상시킴.
      - RFB는 inception module(https://arxiv.org/abs/1409.4842)과 유사한 아이디어를 적용하였으며, 서로 다른 convolution kernels을 가진 여러 branch들을 통해 여러 스케일과 receptive fields의 feature들을 capture하고 최종적으로 이들을 merge함.

1) 다중-스케일 특징 학습(multi-scale feature learning)

- **특징 피라미드(Feature Pyramid)**
  - 통합 특징과 예측 피라미드의 장점을 결합한 방법임.
  - 관련 연구 1-FPN의 원래 연구(https://arxiv.org/abs/1612.03144)
    - 서로 다른 스케일의 feature들을 하향식(top-down fashion)으로 측면 연결(lateral connections)과 통합하여 스케일에 불변하는 feature map을 구축하고, 이러한 특징 피라미드 상에서 여러 스케일에 종속적인 classifiers를 학습시킴.
    - deep layer의 의미적으로 풍부한 feature들은 shallow layer의 공간적으로 풍부한 feature들을 강화시키는데 사용됨.
    - 하향식과 측면 feature들은 element-wise summation 또는 concatenation으로 결합되었으며, 작은 convolutions을 이용하여 차원을 감소시킴.
    - object detection 뿐만 아니라, 다른 응용 분야에서도 상당한 개선을 보여줬으며, 다중 스케일 특징 학습에 있어서 SOTA의 결과를 보임.
    - FPN block(아래 그림 참고)을 변경함으로써, 많은 종류의 변종 FPN들이 개발되어짐.



    - 특징 조합을 위한 일반적인 프레임워크
      - 하향식 features들은 2배 업샘플링되고 상향식 features들과 융합되어짐.
      - 융합 방법은 element-wise sum, multiplication, concatenation 등이 될 수 있음.
      - 의미 정보를 보강하고 메모리 비용을 줄이기 위해 convolution 및 normalization layers가 삽입될 수 있음.

- **1) 다중-스케일 특징 학습(multi-scale feature learning)**
  - **특징 피라미드(Feature Pyramid)**
    - 관련 연구 2-FPN의 변종 연구(RON(https://arxiv.org/abs/1707.01691), RefineDet(https://arxiv.org/pdf/1711.06897.pdf))
      - 측면 연결을 이용하여 스케일에 불변하는 feature map을 구축함.
      - 카테코리별 classifier에 따라 region proposal을 생성하는 FPN과 달리, proposal 생성을 생략함으로써 원래 FPN보다 더욱 더 효율적인 방법임.
    - 관련 연구 3-FPN의 변종 연구(RRC(https://arxiv.org/abs/1704.05776), https://arxiv.org/abs/1705.09587)
      - 서로 다른 layer의 features들 사이의 문맥 정보를 점진적이고 선택적으로 부호화 할 수 있는 새로운 구조를 제안함.
    - 관련 연구 4-FPN의 변종 연구(STDN, http://openaccess.thecvf.com/content_cvpr_2018/papers/Zhou_Scale-Transferrable_Object_Detection_CVPR_2018_paper.pdf)
      - super resolution tasks(https://arxiv.org/abs/1707.02921, https://arxiv.org/abs/1609.05158)로부터 영감을 얻은 방법임.
      - 다중 검출 스케일을 교차하는 스케일 간의 일관성 특성(inter-scale consistency nature)을 명시적으로 탐구한 새로운 transform block을 사용하여 고해상도 feature map을 개발함.

2) 영역 특징 부호화(region feature encoding)

- two-stage detectors에서 영역 특징 부호화는 proposal로부터 features들을 고정된 길이의 feature vector로 추출하기 위한 중요한 단계임.
- 관련 연구 1-cropped region proposals(R-CNN, https://arxiv.org/abs/1311.2524)
  - 전체 이미지로부터 crop된 region을 proposal 하였으며, 이를 위해 crop된 region을 양선형 보간법(bilinear interpolation)을 적용하여 고정된 크기의 patch(224 x 224)로 resize하고 deep convolution feature extractor를 적용함.
  - 고해상도의 region features를 부호화 하였지만, 계산량이 많은 단점을 갖고 있음.
- 관련 연구 2-ROI Pooling layer(Fast R-CNN(https://arxiv.org/abs/1504.08083), Faster R-CNN(https://arxiv.org/abs/1506.01497))
  - 각각의 region을 n x n(기본은 7 x 7) 크기의 cell로 나누고 최대 신호를 가진 neuron만이 feedforward 단계로 진행됨.
  - max pooling과 유사하지만 서로 다른 크기를 가진 region을 이용함.
  - 다운샘플링된 feature map으로부터 feature를 추출하며 그 결과 작은 크기의 object를 다루는데 어려움을 야기함.
- 관련 연구 3-ROI Warping layer(https://arxiv.org/abs/1512.04412)
  - 양선형 보간을 이용하여 region에 대한 feature를 부호화함.
  - DCNN의 다운 샘플링 연산으로 인해, 원 영상에서 object 위치가 잘못 정렬될 수 있으며, ROI Pooling 및 ROI Warping layers가 처리할 수 없는 다운 샘플링된 feature map이 있게 됨.
- 관련 연구 4-ROI Align layer(Mask R-CNN, https://arxiv.org/abs/1703.06870)
  - ROI Warping과 ROI Pooling처럼 grid border를 양자화(quantizing)하는 대신에, 각 grid 내에서 부분적으로 샘플링된 위치에 양선형 보간을 적용하여 양자화 문제를 해결하고자 함.
- 관련 연구 5-Precise ROI Pooling(PrROI Pooling) layer(https://arxiv.org/abs/1807.11590)
  - ROI Align에 기반하고 있음.
  - 좌표에 대한 어떠한 양자화도 수행하지 않으므로, bounding box 좌표에 대한 연속적인 gradient를 갖게 됨.
- 관련 연구 6-Position Sensitive ROI Pooling((PSROI Pooling) layer(R-FCN, https://arxiv.org/abs/1605.06409)
  - 다운 샘플링된 region feature들의 공간 정보를 보강하기 위해, 다운 샘플링된 feature들의 상대적인 공간 정보를 보존함.
  - 생성된 region feature map의 각 channel은 상대적인 공간적 위치에 따라 입력 region의 subset에만 상응하게 됨.
- 관련 연구 7-Feature Selective Network(with PSROI, https://arxiv.org/abs/1711.08879)
  - PSROI Pooling에 기반하고 있음.
  - sub-region과 종횡비 사이의 변이(disparities)를 활용하여 강건한 region features를 학습하고자 함.
  - 제안된 네트워크는 light-weight head에 의한 초기 region features들을 정제할 수 있도록 선택적으로 pooling된 sub-region과 종횡비 정보를 부호화시킴.

- **2) 영역 특징 부호화(region feature encoding)**
  - 관련 연구 8-CoupleNet(https://arxiv.org/abs/1708.02863)
    - ROI Pooling layer와 PSROI Pooling layer에서 생성된 출력을 결합하여 region features들을 추출하는 방식임.
    - ROI Pooling layer
      - global region에 대한 정보를 추출하나, 가려짐이 심한 object들을 다루는데 어려움을 겪음.
    - PSROI Pooling layer
      - ROI Pooling layer와는 달리 local 정보에 보다 많은 초점을 두게 됨.
    - element-wise summation 연산을 이용하여 ROI Pooling 및 PSROI Pooling으로부터 생성된 feature들을 보강시켰으며, 더욱 더 강력한 feature들을 생성하게 됨.
  - 관련 연구 9-Deformable ROI Pooling layer(DCN, https://arxiv.org/abs/1703.06211)
    - 각각의 grid에 대한 offset을 학습시키고 이를 grid의 중심에 추가함으로써 정렬된 ROI Pooling을 일반화시킴.
    - 초기의 region feature를 추출하기 위해 sub-region은 regular한 ROI Pooling layer에서 시작하며, 추출된 feature는 보조(auxiliary) network에 의한 offset를 regression하는데 사용되어짐.
    - 고정된 receptive fields를 이용하는 제약을 부가하지 않음으로써, 자동적으로 이미지의 내용을 모델링 할 수 있음.

- **3) 문맥 추론(contextual reasoning)**
  - 문맥 추론은 object detection에 있어서 중요한 역할을 함.
  - object들은 특정한 환경에서 나타나는 경향이 있으며, 때론 다른 종류의 object들과 함께 존재함.
    - 예: 새는 일반적으로 하늘을 날고 있음.
  - 문맥 정보를 효율적으로 활용함으로써 검출 성능을 개선할 수 있으며, 특히 작은 object나 가려짐(occlusion) 등의 불충분한 cue를 가진 object들을 검출하는데 도움이 됨.
  - object들과 이들을 둘러싸고 있는 문맥 사이의 관계(relationship)를 학습시키는 것은 detector가 시나리오를 이해할 수 있는 능력을 개선하는데 도움이 됨.
  - 전통적인 object detection 알고리즘들에서는 문맥 활용을 위한 연구가 진행(https://vision.cornell.edu/se3/wp-content/uploads/2014/09/context_review08_0.pdf)되었지만, 딥러닝 기반의 object detection에서는 활발한 연구가 수행되지 않았음.
    - convolutional networks는 이미 계층적 특징 표현(hierarchical feature representations)으로부터 문맥 정보를 암시적으로 capture 할 수 있기 때문임.
    - 하지만 여전히 문맥 정보를 활용하기 위한 다양한 연구들이 진행되고 있음.
    - 관련 연구 1(Revisiting RCNN, https://arxiv.org/abs/1803.06799)
      - 때로는 문맥 정보를 활용하는 것이 검출 성능을 저하시킬 수 있다는 것을 보여줌.

- ## 3) 문맥 추론(contextual reasoning)
  - object detection을 위한 문맥 추론 관련 연구들은 크게 전역 문맥 추론(global context reasoning)과 영역 문맥 추론(region context reasoning)의 2가지 양상으로 나뉨.
  - **i) 전역 문맥 추론(global context reasoning)**
    - 전체 이미지 내 문맥으로부터 학습을 수행함.
    - 전통적인 detector들이 이미지 내 특정한 region을 object로 분류하도록 시도하는 것과 달리, 전역 문맥 추론은 특별한 관심 영역을 분류하기 위해 문맥 정보(예: 이미지의 나머지 부분에 대한 정보)를 사용함.
      - 예: 전통적인 detector에서는 이미지에서 야구공을 검출하는 것은 다른 운동에서 사용하는 공과 혼돈을 야기할 수 있기 때문에 어려운 문제임.
      - 이미지의 나머지 부분으로부터 문맥 정보가 사용된다면(예: 야구장, 운동선수, 야구 방망이), 야구공 object를 식별하기가 보다 수월해짐.
    - 관련 연구 1(ION, https://arxiv.org/abs/1512.04143)
      - recurrent neural network을 이용하여 4 방향에서 전체 이미지에 걸쳐 문맥 정보를 부호화시킴.
    - 관련 연구 2(DeepID-Net, https://arxiv.org/abs/1409.3505)
      - 각 이미지에 대한 카테고리별 score를 학습시켰으며 이러한 score는 object 검출 결과와 연결된 문맥적인 feature로서 사용되어짐.
    - 관련 연구 3(Faster R-CNN의 개선 버전, https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf)
      - 전체 이미지에서 embedding된 feature를 추출하고, 검출 결과를 개선 시키기 위해 region feature들과 연계시킴.
    - semantic segmentation을 통하여 전역 문맥 정보를 활용한 연구들
      - 정확한 픽셀 수준의 annotation 때문에, segmentation feature map은 강력한 공간 정보를 capture할 수 있음.
      - 관련 연구 4(Mask R-CNN(https://arxiv.org/abs/1703.06870), https://arxiv.org/abs/1512.04412)
        - 통합된 instance segmentation을 학습시켰으며, pixel 수준의 감독(supervision)을 이용하여 detector를 최적화시킴.
        - 다중 task 최적화로서 detection과 segmentation 목표(objectives)를 함께 최적화시킴.
        - segmentation을 통해 검출 성능을 상당히 개선시킬 수 있지만, pixel 수준의 annotation을 얻는 것은 상당히 소모적인 일임.
      - 관련 연구 5(https://arxiv.org/abs/1803.05858)
        - pseudo segmentation annotation을 이용하여 detector을 최적화시켰으며, 희망적인 결과를 보임.
      - 관련 연구 6(DES, https://arxiv.org/abs/1712.00433)
        - segmentation annotations 없이 segmentation mask를 학습시켜 문맥 정보를 활용하는 방법을 도입함.
        - detection과 segmentation 목표(objectives)를 함께 최적화시켰으며 더욱 더 변별력있는 feature map을 이용하여 원래 feature map을 부유하게 만듬.

- **3) 문맥 추론(contextual reasoning)**
  - **ii) 영역 문맥 추론(region context reasoning)**
    - 영역을 둘러싸고 있는 문맥 정보를 부호화시키며, object와 이들을 둘러싸고 있는 영역 사이의 intersection을 학습시킴.
    - 문맥에 따라 서로 다른 위치와 카테고리별 object의 관계를 직접적으로 모델링하는 것은 상당히 어려움.
    - 관련 연구 1(SMN, https://arxiv.org/abs/1704.04224)
      - 공간 메모리(spatial memory)기반의 모듈을 도입함.
      - 공간 메모리 모듈은 object instance를 pseudo "image" 표현으로 다시 구성함으로써 instance 수준의 문맥을 capture함.
      - 이러한 표현은 추후에 object 관계를 추론하는데 사용되어짐.
    - 관련 연구 2(SIN, https://arxiv.org/abs/1807.00119)
      - 장면 문맥 정보와 object 관계를 고려함으로써 object detection을 그래프 추론 문제로 공식화시킴(formulated).
      - 각각의 object는 그래프의 노드로 취급되며, 서로 다른 object들 사이의 관계는 그래프의 에지로 취급됨.
    - 관련 연구 3(https://arxiv.org/abs/1803.07066)
      - 기존의 region feature 추출 기법들을 통합하여 일반적인 관점을 제공하는 완전히 학습 가능한(fully learnable) object detector를 제안함.
      - ROI Pooling 방법에서 휴리스틱한 선택을 제거하였으며, proposal을 넘어서는 문맥을 포함시켜 가장 중요한 부분을 자동적으로 선택하게 됨.
    - 문맥 정보의 부호화를 위해, region proposal을 둘러싸는 이미지 feature들을 추가하여 region feature들을 암묵적으로 부호화시킨 연구들
      - 관련 연구 4(https://arxiv.org/abs/1505.01749)
        - region proposal의 feature들을 부호화하는 것 이외에도, 원래 object proposal의 다양한 sub-region(경계, 중앙, 문맥 region 등)에서 features들 추출하여 원래 region feature들과 연계시킴.
      - 관련 연구 5(MS-CNN, https://arxiv.org/abs/1607.07155)
        - proposal window의 크기를 확대하고 이러한 features들을 원래 features들과 연계시킴으로써, 국소(local) 문맥을 추출함.
      - 관련 연구 6(GBD-Net, http://www.cs.toronto.edu/~byang/papers/gbd_eccv16.pdf)
        - 다중 스케일의 sub-region으로부터 feature를 추출함.
        - detection을 위해 모든 문맥 정보가 도움이 되는 것은 아니기 때문에, 서로 다른 region 정보의 전송(transmission)을 제어하도록 gated function을 학습시킴.

- **4) 변형 가능한 특징 학습(deformable feature learning)**
  - 우수한 detector는 object의 비정형 변형(nonrigid deformation)에 강건해야 함.
  - 딥러닝 이전
    - 관련 연구 1(DPMs, http://cs.brown.edu/people/pfelzens/papers/lsvm-pami.pdf)
      - deformable coding 기법을 이용하여 object를 여러개의 구성 요소에 해당하는 부분(part)으로 표현함으로써, 비정형 object 변환에 강건한 detector를 만듬.
  - 딥러닝 이후
    - 딥러닝 기반의 detector가 object의 변환을 모델링 하기 위해, object의 부분을 명시적으로(explicitly) 모델링하는 다양한 detection 프레임워크들이 제안됨.
    - 관련 연구 2(DeepID-Net, https://arxiv.org/abs/1412.5661)
      - 서로 다른 object 카테고리를 경유하는 변환 정보를 부호화 하기 위해 변형을 알고 있는(deformable-aware) pooling layer가 제안됨.
    - 관련 연구 3(DCN(https://arxiv.org/abs/1703.06211), Deformable ConvNets v2(https://arxiv.org/abs/1811.11168))
      - feature map의 regular한 샘플링 위치에서 샘플링된 정보를 증강시키기 위해 보조 위치에 대한 offeset을 자동적으로 학습시키는 deformable convolutional layers를 고안함.

Learning Strategy

1. Training Stage

- **object detector를 훈련시키기 위한 학습 전략으로는 크게 데이터 증강(data augmentation), 불균형 샘플링(imbalance sampling), localization 정제(refinement), cascade 학습, 기타(다른 종류의 학습) 전략 등이 있음.**

- **1) 데이터 증강(data augmentation)**
  - 데이터 부족 및 더 많은 훈련 데이터를 이용한 성능 향상 등의 이유로 데이터 증강은 모든 딥러닝 방법들에서 중요한 부분을 차지함.
  - 관련 연구 1(Faster R-CNN, https://arxiv.org/pdf/1506.01497.pdf)
    - 훈련 데이터를 증가시키고 여러개의 시각적 속성을 갖는 patch들을 생성하기 위해 훈련 이미지에 horizontal flips을 적용함.
  - 관련 연구 2(SSD(https://arxiv.org/abs/1512.02325), MS-CNN(https://arxiv.org/abs/1607.07155), SNIPER(https://arxiv.org/abs/1805.09300))
    - one-stage detectors들로써, 보다 다양한 데이터 증강 전략을 사용함.
    - rotation, random crops, expanding, color jittering 등을 적용하였으며, 검출 정확도를 상당히 향상시킬 수 있었음.

- **2) 불균형 샘플링(imbalance sampling)**
  - object detection에서 positive 샘플과 negative 샘플 사이의 불균형은 결정적인 문제임.
  - proposal로 추정된 대부분의 관심 영역은 사실 배경 이미지들이 대부분이며, 이들 중의 극소수만 positive instance(또는 object)를 차지하고 있음.

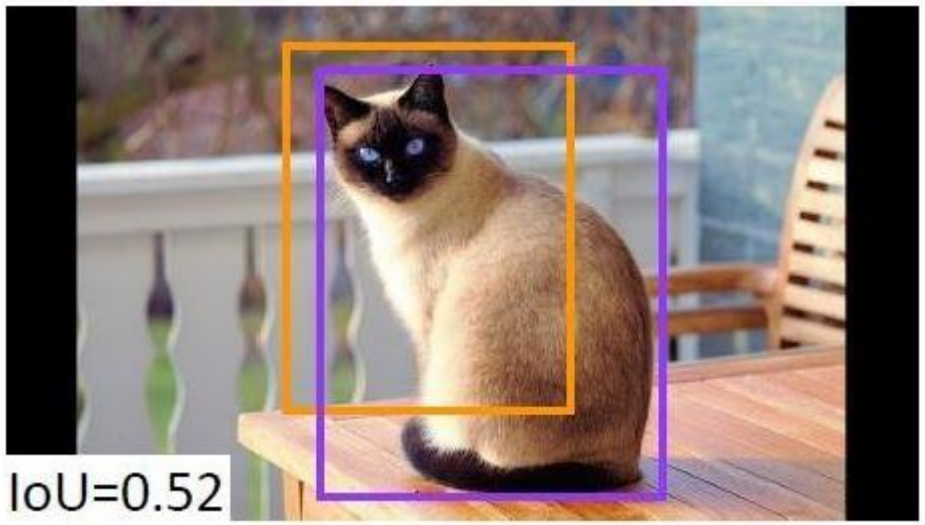
  - detector를 훈련시키는 동안에 불균형 문제를 야기하며 발생할 수 있는 클래스 불균형(class imbalance), 난이도 불균형(difficulty imbalance)의 2가지 문제점이 있음.
  - **i) 클래스 불균형(class imbalance)**
    - 대부분의 후보 proposal들은 배경에 속하며, 단지 몇몇의 proposal만 object들을 포함하고 있기 때문에 발생되는 문제임.
    - 훈련 동안에 배경 proposal이 gradient를 지배하는 결과를 보이게 됨.
  - **ii) 난이도 불균형(difficulty imbalance)**
    - 클래스 불균형 문제와 밀접한 관련이 있음.
    - 클래스 불균형으로 인해 대부분의 배경 proposal을 분류하는 것은 훨씬 쉬워지는 반면에, object를 분류하는 것은 더욱 어려워지게 됨.

- **2) 불균형 샘플링(imbalance sampling)**
  - **클래스의 불균형 문제를 다루기 위한 다양한 방법들**
    - R-CNN과 Fast R-CNN 등의 two-stage detectors들은 대다수의 negative 샘플들을 제거시키고 향 후 분류를 위해 2000개의 proposal만을 남김.
    - 관련 연구 1-random sampling(Fast R-CNN, https://arxiv.org/abs/1504.08083)
      - 2000개의 proposal로부터 random하게 negative 샘플들이 샘플링되고, 클래스 불균형의 악영향을 더욱 줄이기 위해서 mini-batch 시 positive 와 negative 샘플의 비율을 1:3으로 고정시킴.
      - 클래스 불균형 문제를 다룰 수 있긴 하나, negative proposal에 대한 정보를 완벽하게 활용하기는 어려움.
        - 어떤 negative proposal들은 이미지에 대한 풍부한 문맥 정보를 포함하고 있기도 하고, 어떤 hard한 proposal들은 검출 성능을 개선시키는데 도움을 줄 수 있음.
    - 관련 연구 2-hard negative sampling(SSD, https://arxiv.org/abs/1512.02325)
      - random sampling 시 negative proposal 정보를 활용하고자 함.
      - 전경 및 배경의 비율을 고정시키긴 했지만, 모델을 갱신 할 때 가장 어려운 negative proposal들을 샘플링함.
        - classification loss가 더 높은 negative proposals들이 훈련을 위해서 선택되어짐.

  - **난이도 불균형 문제를 다루기 위한 다양한 방법들**
    - loss function을 주의깊게 설계하여 샘플링 하는 방법들이 주를 이룸.
    - 관련 연구 3-focal loss(RetinaNet, https://arxiv.org/abs/1708.02002)
      - 쉬운(easy) 샘플에 대한 신호를 진압하는(suppress) 방법으로서 모든 쉬운 샘플들을 버리는 것 대신에, 각 샘플들에게 loss 값에 대한 중요도 가중치(importance weight)를 할당시킴.
      - 알파와 감마 파라미터를 이용하여 중요도 가중치를 조절하였으며, 쉬운 샘플들의 gradient 신호를 억제시킴으로써, 훈련 과정에서 어려운(hard) proposal에 대해 더욱 더 집중할 수 있게 됨.
    - 관련 연구 4-gradient harmonizing mechanism(GHM, https://arxiv.org/abs/1811.05181)
      - focal loss와 유사한 아이디어를 적용함.
      - 쉬운 proposal을 억제시켰을 뿐만 아니라, outlier에 대한 부정적인 영향을 피하도록 만듬.
    - 관련 연구 5-online hard example mining(OHEM, https://arxiv.org/abs/1604.03540)
      - SSD와 유사한 원리를 적용함.
      - 훈련에 사용되는 hard한 example들을 자동적으로 선택하며, SSD와는 다르게 카테고리별 정보는 무시하고 난이도 정보만을 고려하였음.
      - 따라서, mini-batch 시 전경과 배경의 비율을 고정시키기 않음.
      - object detection task에서는 클래스 불균형보다 어려운 샘플들이 훨씬 더 중요한 역할을 한다고 주장함.

3) localization 정제(refinement)

- object detector는 각 object에 대해 tight한 localization 예측(bounding box 또는 mask)을 제공해야 함.
- localization 개선을 위해 원래 proposal 예측을 정제하기 위한 다양한 연구들이 수행됨.
- 예측(predictions)은 보통 object의 가장 구별되는 부분에 초점을 두기 때문에 반드시 object를 포함한 영역이 아닐 수 있으며, 따라서 정확한 localization이 어려울 수 있음.
- 어떤 시나리오에서는 detection 알고리즘이 높은 품질의 예측(높은 IoU 임계치(threshold))을 하도록 요구함.
- 아래 그림은 높은 IoU 임계치 정책(regime) 때문에 detector가 어떻게 실패하는지에 대해 나타내고 있음.



- 보라색 box: ground truth, 주황색 box: 예측 결과
  - 낮은 IoU가 필요한 시나리오에서는 이러한 예측 결과는 올바르지만, 높은 IoU 임계치를 적용하면 object와 겹쳐짐이 불충분하기 때문에 false positive의 결과를 보이게 됨.
- localization 정제를 위한 일반적인 접근은 고품질의 proposal을 생성하는 것임(**4. Proposal Generation** 참고).

- **3) localization 정제(refinement)**
  - 기본적인 localization 정제를 이용한 연구들
    - 관련 연구 1(R-CNN, https://arxiv.org/abs/1311.2524)
      - localization 정제를 위해 L-2 auxiliary bounding box regressors가 학습되어짐.
    - 관련 연구 2(Fast R-CNN, https://arxiv.org/abs/1504.08083)
      - end-to-end 훈련 계획을 통해 smooth L1 regressors가 학습되어짐.

  - localization 정제를 더 많이 하기 위해 auxiliary models을 학습시킨 연구들
    - 관련 연구 3(MR-CNN, https://arxiv.org/abs/1505.01749)
      - 학습된 예측을 정제하기 위해 R-CNN을 적용한 iterative bounding box regression을 도입함.
      - 예측은 여러번 정제가 됨.
    - 관련 연구 4(LocNet, https://arxiv.org/abs/1511.07763)
      - 각 bounding box의 분포를 모델링하고 학습된 예측을 정제시킴.
    - 관련 연구 3~4 모두 검출 파이프라인에서 별도의 구성 요소가 필요하며, 이는 함께 최적화하는 것을 막게 됨.

  - 수정된 objective functions을 이용하여 통합된 프레임워크를 설계하는데 중점을 둔 연구들
    - 관련 연구 5(Multi-Path Network, https://arxiv.org/abs/1604.02135)
      - 다양한 품질 측정 항목(quality metrics)을 목적으로 하는 integral loss를 이용하여 최적화된 classifiers ensemble을 개발함.
      - 각각의 classifier는 특정한 IoU 임계치에 의해 최적화되며, 최종 예측 결과는 이러한 classifier들로부터 merge가 됨.
    - 관련 연구 6(Fitness-NMS, https://arxiv.org/abs/1711.00164)
      - proposals과 objects들 사이의 IoU에 대해 새로운 fitness score function을 학습시킴.
      - 기존의 존재하는 detector들은 "best"한 예측 대신에 "qualified"한 예측을 찾는 것을 목표로 했으므로, 고품질의 proposal과 저품질의 proposal이 동등한 중요성을 부여받았다고 주장함.
      - 위의 문제를 해결하기 위해 상당히 중복되는 proposals들에 대해서 더 높은 중요성을 부여함.
      - object들의 IoU 예측이 최대치를 갖는 IoU의 상한(upper bound)으로 구성된 set을 기반으로 bounding box regression loss를 도출함.
    - 관련 연구 7(Grid R-CNN, https://arxiv.org/abs/1811.12030)
      - CornerNet과 DeNet에서 영감을 얻어 제안된 방법임.
      - linear bounding box regressor를 corner 기반 메커니즘을 이용하여 corner keypoints를 찾는 원리로 대체시킴.

- **4) Cascade 학습**
  - 보다 강력한 classifiers를 구축하기 위해, cascade 방식을 이용하여 주어진 classifiers의 출력으로부터 정보를 수집하는 coarse-to-fine 방식의 학습 전략을 말함.
  - 딥러닝 이전
    - 관련 연구 1(Viola and Jones Detector, https://www.cs.cmu.edu/~efros/courses/LBMV07/Papers/viola-cvpr-01.pdf)
      - 강건한 face detector를 훈련시키기 위해 cascade 학습 전략을 적용한 첫번째 연구임.
      - 먼저, 경량의 detector가 대다수의 쉬운 negative 샘플들을 제거시키고, detector 훈련을 위해 다음 단계로 보다 어려운 proposal들이 전해지게 됨.

  - 딥러닝 이후
    - 관련 연구 2(CRAFT, https://arxiv.org/abs/1604.03239)
      - cascaded 학습 전략을 이용하여 RPN과 region classifiers를 학습시킴.
      - standard RPN을 먼저 학습시키고, 그 다음에 2가지 class의 Fast R-CNN을 학습시킴으로써 대다수의 쉬운 negative들을 제거함.
      - 나머지 샘플들은 2개의 Fast R-CNN으로 구성되는 cascade region classifiers를 구축하는데 사용되어짐.
    - 관련 연구 3(SDP and CRC, http://www-personal.umich.edu/~wgchoi/SDP-CRC_camready.pdf)
      - 서로 다른 layer 상에서 다른 scale을 갖는 object들을 대응하기 위해 layer-wise cascade classifiers를 도입함.
      - 여러개의 classifiers들이 서로 다른 feature map 상에 배치가 되었으며, shallow layers 내 classifiers들은 쉬운 negative들을 제거시킴.
      - 나머지 샘플들은 classification을 위해 보다 deep한 layers들로 전해지게 됨.
    - 관련 연구 4(RefineDet(https://arxiv.org/abs/1711.06897), Cascade R-CNN(https://arxiv.org/abs/1712.00726))
      - object의 localization을 정제하는데 cascade 학습 방법을 적용함.
      - 여러 단계의 bounding box regressors를 구축하였고, 각 단계마다 서로 다른 품질 측정 방식(quality metrics)을 이용하여 훈련시킴으로서 bounding box 예측을 정제시킴.
    - 관련 연구 5(Revisiting RCNN, https://arxiv.org/abs/1803.06799)
      - Faster R-CNN이 실패하는 경우를 분석하였고, object의 localization이 좋다고 할지라도, 몇몇의 classification error가 발생한다는 것을 확인함.
      - classification과 regression을 위한 공동(joint)의 multi-task 최적화와 feature의 공유로 인해 차선(sub-optimal)의 특징 표현(feature representation)밖에 하지 못한 것에 원인이 있다고 주장함.
      - 또한, Faster R-CNN의 큰 receptive field는 detection 과정에서 너무 많은 noise를 만들게 된다고 주장함.
        - vanilla R-CNN이 이러한 문제에 대해 강건하다는 것을 발견함.
      - 최종적으로, 상호 보완을 위한 Faster R-CNN과 R-CNN 기반의 cascade detection 시스템을 구축함.
        - Faster R-CNN으로부터 잘 훈련된 초기의 예측 set을 얻은 후, 이러한 예측은 결과를 보다 정제하기 위한 용도로 R-CNN을 훈련시키는데 사용함.

- **5) 기타 전략(다른 종류의 학습 전략)**
  - 관심이 가는 다른 종류의 학습 전략들이 있지만, 활발하게 연구되지는 않음.
  - 이러한 학습 전략은 크게 적대적 학습(adversarial learning), 밑바닥부터 훈련하기(training from scratch), 지식 증류(knowledge distillation) 등이 있음.

  - **i) 적대적 학습(adversarial learning)**
    - 생성 모델(generative models)에 있어서 적대적 학습은 상당한 진보를 보여주었음.
    - 적대적 학습을 적용한 가장 유명한 연구는 GAN(https://arxiv.org/abs/1406.2661)임.
      - 생성자(generator)와 구별자(discriminator)가 경쟁을 하는 구조임.
        - 생성자는 noise vector를 입력으로 이용하여 fake 이미지 생성을 위한 데이터 분포를 모델링하려 하며, 이러한 fake 이미지들은 구별자를 혼란시키는데 사용이 됨.
        - 반면에, 구별자는 fake 이미지로부터 real 이미지를 식별하기 위해 생성자와 경쟁을 하게 됨.
    - GAN과 다양한 종류의 변종들은 여러 domain에서 효율성을 보였으며, object detection에서도 적용된 사례가 있음.

    - object detection에 GAN을 적용한 연구들
      - 관련 연구 1(Perceptual GAN, https://arxiv.org/abs/1706.05274)
        - 작은 object 검출을 위해 제안된 새로운 프레임워크임.
        - 학습 가능한(learnable) 생성자는 적대적 계획(adversarial scheme)을 통해 작은 object의 고해상도용 특징 표현을 학습시킴.
        - 생성자는 저해상도의 작은 region feature들을 고해상도의 feature들로 전이(transfer)하기 위해 학습되었으며, 실제 고해상도 feature들을 식별한 구별자와 경쟁을 하게 됨.
        - 최종적으로 생성자는 작은 object들을 고품질의 feature를 생성하기 위한 방법을 배우게 됨.
      - 관련 연구 2(A-Fast-R-CNN, https://arxiv.org/abs/1704.03414)
        - 생성된 적대적 example들을 이용하여 훈련된 방법임.
        - 어려운 샘플들은 long tail에 있다고 주장하였으며, 가려짐과 변형을 자동으로 생성하는 2개의 새로운 block을 도입시킴.
        - 학습된 mask는 region classifiers를 따르는 region feature 상에서 생성되었으며, detector는 더 많은 적대적 example들을 수락하게 되므로, 더욱 강건해질 수 있음.

- **5) 기타 전략(다른 종류의 학습 전략)**
  - **ii) 밑바닥부터 훈련하기(training from scratch)**
    - 최신의 object detector들은 ImageNet을 이용하여 pre-trained된 classification models에 상당히 의존적임.
    - 하지만, loss function의 편향(bias)과 classification과 detection 사이의 데이터 분포는 성능에 적대적 영향을 줌.
    - detection task에 대한 Finetuning은 이러한 문제를 경감시켜주지만, 편향을 완벽하게 제거할 수는 없음.
    - 또한, 새로운 domain에서 detection을 위해 classification model을 전이하는 것은 더욱 더 도전적인 일임.
      - 예: RGB 데이터를 MRI 데이터로 전이하는 것
    - 이러한 이유로 pretrained된 모델에 의존하는 대신에 밑바닥(scratch)부터 detector를 훈련시킬 필요성이 있게 됨.
      - 이 때의 어려움은 object detection을 위한 훈련 데이터가 충분하지 않다는 것이며, 과적합(overfitting)을 야기시킬 수 있다는 것임.
      - image classification과는 다르게, object detection은 bounding box 수준의 annotation이 필요하며, 대규모의 dataset을 annotation하는 것은 많은 시간과 노력이 필요하게 됨.
        - 예: image classification을 위한 ImageNet의 경우 1000개의 카테고리를 가지고 있지만, 그 중 200개만이 annotation을 가지고 있음.
      - 관련 연구 1(DSOD, https://arxiv.org/abs/1708.01241)
        - densely하게 연결된 network 구조를 이용하여 deep한 supervision을 하는 것은 최적화에 대한 어려움을 상당히 줄일 수 있다고 주장함.
      - 관련 연구 2(GRF-DSOD, https://www.researchgate.net/publication/321511551_Learning_Object_Detectors_from_Scratch_with_Gated_Recurrent_Feature_Pyramids)
        - DSOD에 기반한 방법이며, DSOD보다 더욱 강력한 방법임.
        - 서로 다른 스케일을 가지는 object들을 위해 중간(intermediate) layer의 supervision 강도(intensities)를 동적으로 조절하는 gated recurrent feature pyramid를 제안함.
        - 공간 및 의미 정보를 하나의 예측 layer로 압축(squeeze)시키기 위해 recurrent한 feature pyramid를 정의하였으며, 더욱 더 빠른 수렴이 가능하도록 파라미터의 수를 줄임.
        - feature pyramid의 gate-control 구조를 이용해 object들의 크기에 기반한 서로 다른 스케일에서의 supervision을 적응적으로 조정하게 됨. ```
      - 관련 연구 3(https://arxiv.org/abs/1811.08883)
        - MSCOCO를 이용하여 detector를 밑바닥부터 훈련시키는 것은 어렵다는 것을 확인함.
        - vanilla detectors는 적어도 10000장의 annotation된 이미지들을 사용해야만 경쟁력 있는 성능을 얻을 수 있다는 것을 발견함.
        - 이전 연구들과 상반되게 밑바닥부터 훈련시에는 특정한 구조가 필요하지 않다는 것을 증명함.

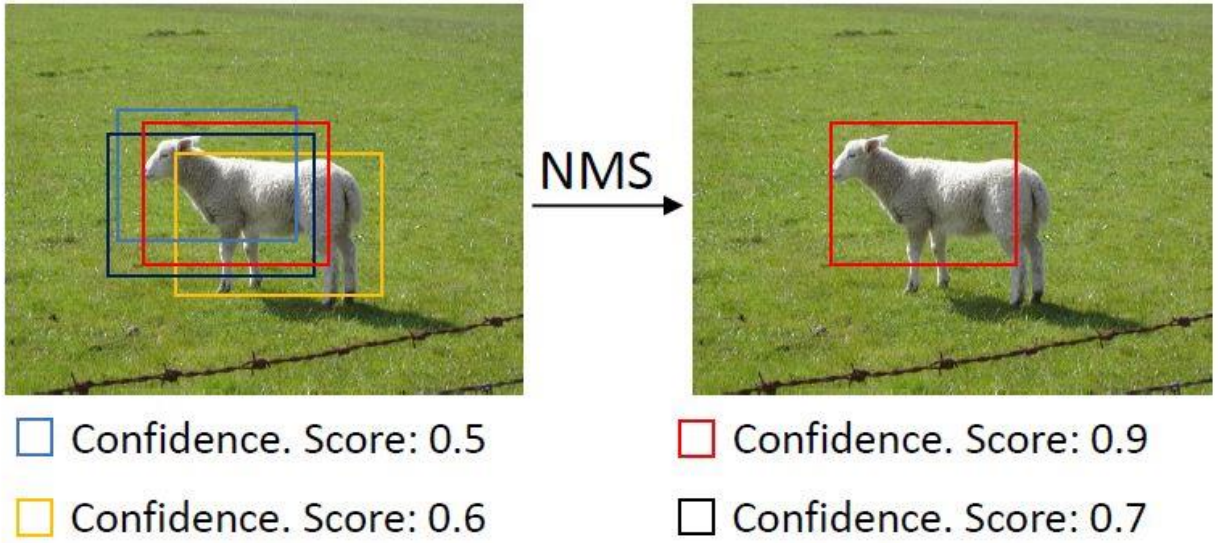
- **5) 기타 전략(다른 종류의 학습 전략)**
  - **iii) 지식 증류(knowledge distillation)**
    - teacher-student 훈련 계획을 이용하여 앙상블(ensemble)된 모델들의 지식을 단일 모델로 증류(distill)하는 훈련 전략임.
    - image classification task에서 처음으로 사용되어짐(https://arxiv.org/abs/1503.02531).

    - object detection에서 지식 증류를 적용한 연구들
      - 관련 연구 1(http://openaccess.thecvf.com/content_cvpr_2017/papers/Li_Mimicking_Very_Efficient_CVPR_2017_paper.pdf)
        - 무겁지만 강력한 detector에 의해 유도되어(guide) 신중하게 최적화가 수행된 경량의 detector가 제안됨.
        - 무거운 detector로부터 지식을 증류함으로써 비교할만한 검출 정확도를 달성하였으며, 빠른 추론 속도를 보임.
      - 관련 연구 2(Revisiting RCNN, https://arxiv.org/abs/1803.06799)
        - teacher-student 훈련 계획을 통해 최적화된 Faster R-CNN 기반의 detector임.
        - R-CNN 모델을 훈련 과정을 유도하기 위한 teacher network로 사용함.
        - 전통적인 단일 모델 최적화 전략들과 비교했을 때, 향상된 검출 정확도를 보임.

2. Test Stage

- object detection task는 dense한 set의 예측을 만들므로, 상당한 중복을 포함하고 있기 때문에 평가 시 직접적으로 예측을 사용할 수 없음.
- 추가로, 검출 성능을 보다 개선하기 위해 몇몇의 다른 학습 전략들을 필요로 함.
- 이러한 전략들을 통해 예측의 품질을 개선시킬 수 있으며, 추론 속도를 빠르게 할 수 있음.

- 테스트 단계에서 활용할 수 있는 전략들로는 중복 제거(duplicate removal), 모델 가속(model acceleration), 기타 효율적인 기법들이 있음.
- **1) 중복 제거(duplicate removal)**
  - object detection 알고리즘은 중복된 예측들이 있는 dense한 예측 set을 만듬.
  - NMS(Non maximum suppression)는 아래 그림과 같이 중복된 false positive 예측을 제거하기 위한 필수적인 요소임.



  - 가장 confidence scores가 높은 box만 남기고, 이를 둘러싸고 있는 다른 boxes들은 제거함으로써 중복된 예측을 제거함.

- **1) 중복 제거(duplicate removal)**
  - SSD(https://arxiv.org/abs/1512.02325)나 DSSD(https://arxiv.org/abs/1701.06659)와 같은 one-stage detection 알고리즘들은 dense한 set으로 구성된 후보 proposal을 생성함.
    - 동일한 object를 둘러싸고 있는 proposal들은 유사한 confidence scores를 갖게 되며, false positive를 야기시킴.
  - two-stage detection 알고리즘들은 sparse한 set으로 구성되는 후보 proposal을 생성함.
    - bounding box regressors가 proposal을 동일한 object로 가깝게 당기므로(pull) 위의 동일한 문제를 야기시킴.
  - 중복된 예측은 false positive로 간주되며, 평가 시 페널티를 받기 때문에 NMS로 이러한 중복된 예측을 제거하는 것이 필요함.

  - 일반적인 NMS 수행을 위한 절차
    - 예측 boxes들을 confidence score에 따라서 정렬하고, 가장 높은 score를 갖는 box를 선택함.
    - 즉, 가장 높은 score를 갖는 box와 겹치는 다른 boxes들간의 IoU 값이 미리 정의된 임계치보다 높으면 score를 0으로 설정하여 제거시킴.
    - 이러한 과정은 모든 남아 있는 예측 boxes들에 대해서 반복적으로 수행됨.
    - IoU 값이 미리 정의된 임계치보다 높다는 이유로, 실제 object를 놓쳐버리는 결과를 보일 수 있으며, clustered object detection에서 상당히 자주 발생하게 됨.

  - NMS 개선을 위한 연구들
    - 관련 연구 1(Soft-NMS, https://arxiv.org/abs/1704.04503)
      - 가장 높은 score를 갖는 box를 제외한 다른 boxes들을 IoU 값에 임계값을 적용하여 바로 제거하는 것이 아니라, confidence score를 연속 함수(선형 또는 가우시안 함수)를 적용하여 서서히 줄이는(decay) 방법을 적용함.
      - clustered object들의 예측을 제거하는 현상을 피할 수 있으며, 많은 공용 벤치마크에서 개선된 결과를 보여줌.
    - 관련 연구 2(Learning NMS, https://arxiv.org/abs/1705.02950)
      - confidence scores와 bounding boxes를 기반으로 NMS를 수행하는 네트워크 아키텍처를 제안하였으며, 감독 방식으로 detector를 훈련시키는 것과는 별개로 최적화가 수행됨.
      - 중복 예측의 이유는 detector가 단 한개의 높은 score를 보상하는(reward) 대신에, object 마다 여러개의 높은 score를 갖는 detection을 의도적으로 권장(encourage)했기 때문이라고 주장함.
      - 이를 바탕으로 아래와 같은 2가지 동기에 따라서 네트워크를 설계함.
        - i) 이중 detection의 경우 페널티를 부가하는 loss를 적용하여 detector가 object 당 하나의 detection 결과를 갖도록 함.
        - ii) object가 두번 이상 검출되는지에 대한 여부를 detector 정보로 제공하기 위해 가까운 detection들을 함께 처리함.
      - detection을 버리는 것 대신에 이미 검출된 object를 포괄하는 detection score를 줄이기 위해 re-scoring task를 이용하여 NMS를 재구성함.

- **2) 모델 가속(acceleration)**
  - 실세계의 object detection application들은 알고리즘의 효율적인 처리를 필요로 하므로 효율성 측정 방식(efficiency metrics)을 고려한 detector의 평가가 중요함.
  - 비록 SOTA 알고리즘들이 공용 데이터셋에서 상당히 우수한 결과를 보이지만, 추론 속도가 느리기 때문에 실세계 applications에 적용하는 것이 어려움.
  - two-stage detector는 proposal generation과 region classification의 2개의 stage를 갖으므로 one-stage detector에 비해서 느림.
    - proposal generation과 region classification을 위해 하나의 네트워크를 이용하는 one-stage detector에 비해서 계산적으로 더 많은 시간을 소비하게 됨.

  - 관련 연구 1(R-FCN, https://arxiv.org/abs/1605.06409)
    - 공간적으로 민감한(spatially-sensitive) feature map을 구축하고, 계산 비용을 공유하기 위해 위치에 민감한(position sensitive) ROI Pooling을 이용하여 feature를 추출함.
    - 카테고리의 개수가 증가함에 따라 공간적으로 민감한 feature map들의 채널 개수는 크게 증가하는 단점이 있음.
  - 관련 연구 2(Light Head R-CNN, https://arxiv.org/abs/1711.07264)
    - 모든 계산을 공유하는 대신에 최종 feature map의 채널 개수를 1024에서 16개로 상당히 줄임.
    - 모든 region에 대해 계산이 공유되진 않지만, 계산 비용을 무시할 수 있게 됨.

  - backbone 측면에서 object detection 시 가장 많은 계산 비용을 차지하는 부분은 특징 추출(feature extraction)임.
  - detection 속도를 가속하는 가장 단순한 아이디어는 detection backbone을 보다 효율적인 backbone으로 대체하는 것임.
    - 관련 연구 3(MobileNet, https://arxiv.org/abs/1704.04861?source=post_page----------)
      - depth-wise convolution layers를 이용하는 효율적인 CNN 모델로서, Tiny SSD(https://arxiv.org/abs/1802.06488), Tiny-DSOD(https://arxiv.org/abs/1807.11013) 등의 연구에 채택되어짐.
    - 관련 연구 4(PVANet, https://arxiv.org/abs/1608.08021)
      - 비선형 계산을 줄이고 추론 속도를 가속하기 위해 CReLu(https://arxiv.org/abs/1603.05201) layer를 적용한 새로운 네트워크 아키텍처를 제안함.
  - 모델 압축 및 양자화 등을 이용하여 모델을 오프라인으로 최적화하는 연구들
    - 관련 연구 5
      - https://arxiv.org/abs/1511.06530
      - https://arxiv.org/abs/1707.06168
      - https://arxiv.org/abs/1412.6115
      - https://arxiv.org/abs/1712.01887
      - https://arxiv.org/abs/1512.06473
      - https://arxiv.org/abs/1510.00149
      - https://arxiv.org/abs/1506.02626
    - 관련 연구 6(TensorRT, https://developer.nvidia.com/tensorrt)
      - 효율적 사용을 위해 학습된 모델의 계산을 최적화함으로써, 추론 속도를 가속화함.

- **3) 기타 학습 전략**
  - test 단계에서 사용할 수 있는 다른 학습 전략들은 검출 정확도를 향상시키기 위해 입력 이미지를 변환(transformation)시키는 것임.

  - 이미지 피라미드는 검출 성능 개선을 위해 널리 사용되고 있음.
    - 서로 다른 스케일에 대한 계층적 이미지 set을 구축하여, 이러한 이미지 모두에 대해서 예측을 수행하고 최종 검출 결과는 각 이미지의 예측을 merge시킴.
    - 관련 연구 1(S3FD(http://openaccess.thecvf.com/content_ICCV_2017/papers/Zhang_S3FD_Single_Shot_ICCV_2017_paper.pdf), https://arxiv.org/abs/1711.06897)
      - 서로 다른 스케일을 가진 object들을 다루기 위해 더욱 더 광범위한 이미지 피라미드 구조를 사용함.
      - 서로 다른 스케일로 test 이미지를 resize하였으며, 각 스케일은 특정한 스케일 범위를 가지는 objects들을 대응하게 됨.
  - Horizontal Flipping을 이용한 연구들
    - 관련 연구 2(Mask R-CNN(https://arxiv.org/abs/1703.06870), https://arxiv.org/abs/1711.06897)
      - test 이미지에 적용하여, 성능을 개선시킴.
  - 이미지 피라미드나 Horizontal Flipping을 이용한 학습 전략은 다양한 스케일을 갖는 object들을 다룸으로써 detector의 성능을 크게 향상시켰지만, 계산 비용 역시 증가하기 때문에 실세계 application에는 적합하지 않음.

Application & Detection Benchmarks

Application

1. Face Detection

- generic object detection과 달리 실세계 application들은 보통 자신만의 구체적인 특성이 있으므로 주의 깊게 고안된 detection 알고리즘을 필요로 함.
- 이미지에서 얼굴을 검출하는 고전적인 컴퓨터 비전 문제이며, 얼굴 검증(verification), 정렬(alignment), 인식(recognition)과 같은 실세계 application을 위한 첫번째 단계에 해당됨.
- generic object detection과 face detection은 중요한 몇가지 차이점이 있음.
  - i) face detection에서 object의 스케일 범위는 generic object detection의 object들보다 훨씬 더 크며, 가려짐이나 blur도 더 많이 발생함.
  - ii) face object는 훨씬 더 강력한 구조적인 정보를 포함하고 있으며, face detection에는 단지 하나의 target 카테고리만 있으면 됨.
  - 위와 같은 특성을 고려해봤을 때, face detection을 개선하기 위해 이용할 수 있는 몇가지 prior들이 있을 수 있으므로, generic object detection을 직접적으로 적용하는 것은 최적의 해결책이 아님.
- 딥러닝 이전
  - sliding window에 기반한 연구들이 주를 이루며, dense한 이미지 grid를 hand crafted features를 이용하여 부호화시킨 후 classifier로 훈련시켜 object의 위치를 찾음.
  - 관련 연구 1(Viola and Jones Face Detector, http://www.face-rec.org/algorithms/Boosting-Ensemble/16981346.pdf)
    - Haar feature와 AdaBoost를 사용하는 cascaded classifiers를 제안하였으며, 실시간 예측 속도와 우수한 성능을 보임.
- 딥러닝 이후
  - image classification에 있어서 딥러닝의 발전 이후 face detection에 있어서도 딥러닝을 기반으로 한 방법들의 성능이 전통적인 방법들을 크게 능가하는 성능을 보이게 됨.
  - Fast R-CNN이나 SSD와 같은 generic detection 프레임워크로부터 확장되었음.
  - 강건한 특징 표현을 학습시키는 것에 초점을 두고 있으며, 상당한 스케일 변화를 다루기 위해 이전에 설명한 다중 스케일 특징 학습 기법들이 널리 사용되었음.

  - **다중 스케일 특징 학습을 이용한 연구들**
    - 관련 연구 2(https://arxiv.org/abs/1701.08289)
      - Fast R-CNN 기반의 프레임워크로서 예측을 위해 다중 스테일 특징을 통합함.
      - 사람의 얼굴 영역은 직사각형 보다는 타원에 가깝기 때문에 검출 결과 bounding boxes를 타원으로 변환시킴.
    - 관련 연구 3(S3FD, http://openaccess.thecvf.com/content_ICCV_2017/papers/Zhang_S3FD_Single_Shot_ICCV_2017_paper.pdf)
      - 큰 범위의 스케일에서 얼굴을 검출하기 위해 서로 다른 feature map 상에서 얼굴을 찾는 one-stage의 접근임.
      - 작은 스케일의 얼굴 정보를 capture하기 위해 보다 큰 feature map에서 예측을 수행함.
      - 실험적 receptive fields에 따라 anchor set을 주의 깊게 설계하였으므로 얼굴을 좀 더 우수하게 matching시킴.

- 딥러닝 이후
  - **다중 스케일 특징 학습을 이용한 연구들**
    - 관련 연구 4(FANet, https://arxiv.org/pdf/1712.00721.pdf)
      - 서로 다른 stage에서 다중 스케일의 feature들을 capture하기 위한 새로운 네트워크 구조를 제안함.
      - 계층적 방식으로 서로 다른 스케일에서 통합된 feature들을 합치는 구조를 가지며, 훈련의 어려움을 줄이기 위해 계층적 loss를 제안함.
    - 관련 연구 5(SSH, https://arxiv.org/abs/1708.03979)
      - 다른 종류의 one-stage detector로서 예측을 위해 서로 다른 스케일 features들을 결합시킴.
    - 관련 연구 6(https://arxiv.org/abs/1612.04402)
      - 작은 얼굴 검출을 위한 상세 분석을 수행하고 여러개의 RPN들로 구성되는 경량의 face detector를 제안함.
      - 각각의 RPN들은 특정한 범위의 스케일을 대응하게 됨.
      - 스케일 변화를 효율적을 다룰 수 있으나 실세계 적용을 위해서는 느린 속도를 보임.
    - 관련 연구 7(SAFD, https://arxiv.org/abs/1706.09876)
      - 상당한 계산 비용 없이 스케일 문제를 해결하는 방법을 제안함.
      - 주어진 이미지 내 얼굴의 스케일 분포를 모델링하는 네트워크를 학습시켰으며, 얼굴이 원하는 스케일에 있는지 확인하기 위해 zoom-in 또는 zoom-out 연산으로 안내(guide)를 함.
      - resize된 이미지는 단일 스케일의 경량 face detector의 입력으로 사용됨.
    - 관련 연구 8(Face R-CNN, https://arxiv.org/abs/1706.01061)
      - RetinaNet(https://arxiv.org/abs/1708.02002)을 기반으로 하였으며, 큰 범위의 스케일에서 얼굴을 다루기 위해 더욱 더 dense한 anchor들을 사용함.
      - 문맥 정보를 계산하고 변별력있는 feature들을 강조하기 위해 attention function을 제안함.
    - 관련 연구 9(MTCNN, https://arxiv.org/abs/1604.02878)
      - coarse-to-fine 방식으로 얼굴을 예측하기 위해 주의 깊게 설계된 3단계의 CNN 모델을 제안함.
      - 또한, 결과를 개선하기 위해 새로운 online hard negative mining을 제안함.
    - 관련 연구 10(Face-MagNet, https://arxiv.org/abs/1803.05258)
      - 보다 정교한 얼굴 표현을 구축하기 위해 RPN과 ROI Pooling 앞에 deconvolution layers set을 배치함으로써 어떠한 skip connections 없이도 작은 얼굴에 대한 정보의 흐름을 허용하도록 함.

- 딥러닝 이후
  - **문맥 정보를 활용한 연구들**
    - face object는 주변의 문맥(일반적으로 사람의 몸에서 나타남)과 강한 물리적인 관계를 가지므로, 문맥 정보를 부호화시켜 검출 정확도를 효율적으로 개선시키기 위한 연구들이 수행됨.
    - 관련 연구 11(FDNet, https://arxiv.org/abs/1802.02142)
      - 이미지 문맥을 capture하기 위해 보다 큰 deformable convolutional kernels을 가진 ResNet 기반의 기법을 제안함.
    - 관련 연구 12(CMS-RCNN, https://arxiv.org/abs/1606.05413)
      - 다양한 범위의 스케일에 대한 얼굴을 다루기 위해 다중 스케일 정보가 region proposal과 ROI detection 모두에서 grouping됨.
      - detector 훈련 시 얼굴 주변의 문맥 정보가 고려됨.
    - 관련 연구 13(PyramidBox, https://arxiv.org/abs/1803.07737)
      - 검출에 어려운 얼굴 문제를 다루기 위해 최신의 context assisted single shot face detector를 제안함.
      - 문맥의 중요성을 관찰하여, 다음의 3가지 측면에서 문맥 정보의 활용을 향상시킴.
        - i) semi supervised 방법으로 고수준의 문맥적인 feature 학습을 supervise하기 위해 새로운 문맥 anchor를 설계함.
        - ii) 적절한 고수준의 문맥 의미(semantic) feature들과 저수준의 얼굴 feature들을 함께 결합하기 위해 저수준의 Feature Pyramid Network를 개발하였으며, single shot으로 모든 스케일에서 얼굴을 예측하도록 함.
        - iii) 최종 출력의 개선을 위한 예측 네트워크의 수용력(capacity)을 증가시키기 위해 문맥에 민감한 구조를 도입함.
      - 서로 다른 스케일에 대한 훈련 샘플들을 증강시키기 위해 data-anchor-sampling 기법을 사용하였으며, 이를 통해 작은 얼굴들에 대한 훈련 데이터의 다양성을 증가시킬 수 있었음.
    - 관련 연구 14(https://arxiv.org/abs/1805.03363)
      - 이미지 문맥을 활용하기 위해 context pyramid maxout 메커니즘을 도입하고 cascaded 방식으로 anchor 기반 detector를 최적화시킨 효율적인 anchor 기반 cascaded 프레임워크를 고안함.
    - 관련 연구 15(https://ai.tencent.com/ailab/media/publications/Detecting_Faces_Using_Inside_Cascaded_Contextual_CNN.pdf)
      - 몸의 부분 정보를 적응적으로 capture하기 위해 two-stream의 contextual CNN을 제안함.
      - shallow layer에서는 얼굴이 아닌 쉬운 영역을 필터링하였으며, 보다 deep한 layer에서는 어려운 샘플들을 필터링함.

- 딥러닝 이후
  - **기타 다양한 연구들**
    - 관련 연구 16(Face R-CNN, https://arxiv.org/abs/1706.01061)
      - loss function의 설계 관점에서 제안된 프레임워크임.
      - vanilla Faster R-CNN에 기반하였으며, 원래 softmax loss를 center loos로 교체함으로써 detector가 얼굴 검출 시 발생하는 large intra-class variance를 줄일 수 있도록 함.
      - 고정된 비율의 online hard negative mining, 다중 스케일 훈련, 다중 스케일 테스트 등과 같이 Faster R-CNN을 개선시키기 위한 여러가지 기법들을 적용하였으며, vanilla Faster R-CNN을 face detection에 적합하도록 만듬.
    - 관련 연구 17(Face R-FCN, https://arxiv.org/abs/1709.05256)
      - vanilla R-FCN에 기반하고 있음.
      - 다양한 얼굴 부분의 기여(contribution)를 구별하고, 최종 score map의 응답을 re-weight하기 위해 위치에 민감한(position-sensitive) average pooling을 새롭게 도입함.
      - FDDB와 WIDER FACE와 같은 공용 벤치마크에서 SOTA의 결과를 보임.

2. Pedestrian Detection

- 지능형 비디오 감시 시스템 등에서 보행자 검출은 본질적이며 중요한 task임.

- 보행자 검출은 generic object detection과는 다른 몇몇의 특성들이 있음.
  - i) 보행자 object들은 종횡비가 약 1.5 정도로 고정되어 잘 구조화된 object들이지만, 커다란 범위의 스케일을 가지고 있음.
  - ii) 보행자 검출은 실세계 application으로써 복잡함(crowding), 가려짐, blurring 등이 자주 나타나므로 도전적인 task임.
    - 예: CityPersons 데이터셋의 특성
      - validation subset에는 전체 3157개의 보행자 annotations이 있으며, 그중 48.8%가 IoU 0.1 이상을 가진 다른 annotated된 보행자들과 겹쳐져 있음.
      - 전체 보행자의 26.4%가 IoU 0.3 이상을 가진 다른 annotated된 보행자들과 상당히 겹쳐져 있음.
    - 상당히 빈번한 복잡한 가려짐은 보행자 검출의 성능에 악영향을 주게 됨.
  - iii) 복잡한 상황으로 인해 교통 신호등, 우체통 등과 같은 더욱 어려운 negative sample들이 있음.

- 딥러닝 이전
  - Viola Jones 프레임워크(http://www.face-rec.org/algorithms/Boosting-Ensemble/16981346.pdf)를 기반으로 확장되었음.
  - object의 위치를 찾기 위해 sliding window 전략을 이용한 Integral Channel Features를 활용하였으며, SVM과 같은 region classifier를 적용함.
  - 초기 연구들은 분류를 위한 강건한 feature descriptors를 고안하는데 초점을 두었음.
  - 관련 연구 1(HOG, https://lear.inrialpes.fr/people/triggs/pubs/Dalal-cvpr05.pdf)
  - 관련 연구 2(https://www.researchgate.net/publication/3766402_General_framework_for_object_detection)
    - 저수준의 visual cue들과 spatial pooling features에 기반한 feature descriptors를 제안함.
  - 보행자 검출 벤치마크에 대해서 희망적인 결과를 보여주었지만, hand-crafted features에 기반하고 있음.

- 딥러닝 이후
  - 딥러닝 기반의 보행자 검출은 뛰어난 성능을 보여주었으며, 공용 벤치마크에서 SOTA의 결과를 보임.
  - 관련 연구 3(https://static.googleusercontent.com/media/research.google.com/ko//pubs/archive/43850.pdf)
    - cascade된 deep convolutional networks를 적용하였으며, tiny model에 의해 대다수의 쉬운 negative 샘플들이 제거되며, 남아 있는 어려운 proposal들은 큰 deep network에 의해서 분류가 됨.
  - 관련 연구 4(https://arxiv.org/abs/1607.07032)
    - decision tree 기반의 프레임워크를 제안하였으며, 다중 스케일의 feature map이 보행자 features들 추출하는데 사용되었고 추후 분류를 위한 boosted decision trees의 입력으로 전해짐.
    - hard negative samples mining을 위해 boosted decision trees에 bootstrapping 전략을 적용하였고 FC layers와 비교했을 때 보다 우수한 성능을 보임.
  - 관련 연구 5(SAF R-CNN, https://arxiv.org/abs/1510.08160)
    - 전체 detection 프레임워크에 여러개의 built-in된 network들을 삽입하였으며, 서로 다른 sub-net를 이용하여 서로 다른 스케일의 보행자 instance들을 검출함.
  - 관련 연구 6(SDP-CRC, http://www-personal.umich.edu/~wgchoi/SDP-CRC_camready.pdf)
    - 보행자 스케일 문제를 다루기 위해 스케일에 종속적인(Scale Dependent) Pooling과 cascade 방식의 rejection을 위한(Cascaded Rejection) Classifiers를 Fast R-CNN에 삽입시킴.
    - 보행자의 높이에 따라, SDP는 적합한 feature map으로부터 region features들을 추출하며, CRC는 shallower layers에서 쉬운 negative 샘플들을 제거시킴.
  - 관련 연구 7(Repulsion Loss, https://arxiv.org/abs/1711.07752)
    - 복잡한 환경에서 보행자 검출을 할 때는 NMS 임계치에 따라 false positives를 많이 만들거나 object를 놓칠 수 있으므로 NMS 임계치가 상당히 민감하게 작용함.
    - 새롭게 제안된 repulsion loss는 proposals를 target objects 속으로 밀어 넣었을뿐만 아니라, 다른 object들과 target proposal들로부터도 멀어지게 함.
  - 관련 연구 8(OR-CNN, https://arxiv.org/abs/1807.08407)
    - 새롭게 제안한 aggression loss에 의해 최적화가 수행되며, proposal들을 object들과 가깝게 만들어줌.
  - 관련 연구 9(https://arxiv.org/abs/1705.02757)
    - 추가적인 features들을 보행자 detector에 적절히 통합시키면, 검출의 정확도를 향상시킬 수 있다고 주장함.
    - 정확도 개선을 위해 유용한 여러 종류의 추가적인 features들에 대해 연구하였으며, 이러한 features들을 활용하는 새로운 방법을 제안함.
    - 새롭게 제안한 component인 HyperLearner는 공동적인 최적화 방식을 통해 추가적인 features들을 vanilla DCNN detector와 통합시켰으며, 추론 단계에서는 추가적인 입력이 필요하지 않음.

- 딥러닝 이후
  - 가려짐을 다루기 위한 연구들
    - 보행자 검출에 있어서, 가장 도전적인 문제 중 하나는 가려짐을 다루는 것임.
    - 이를 위한 직관적인 접근은 part 기반의 모델을 사용하여 일련의 part detector들을 학습시키고 object의 위치를 찾고 분류하기 위해 part detector의 결과를 통합시키는 것임.
    - 관련 연구 10(DeepParts, http://personal.ie.cuhk.edu.hk/~pluo/pdf/tianLWTiccv15.pdf)
      - 여러개의 part detector들로 구성되어 있으며, 신체의 모든 스케일 part를 cover할 수 있는 part pool로부터 중요한 신체 part가 자동적으로 선택되어 훈련되며, 각각의 선택된 part에 대해 detector가 가려짐을 다룰 수 있도록 훈련되어짐.
    - 관련 연구 11(http://www.ee.cuhk.edu.hk/~xgwang/papers/ouyangWcvpr12.pdf)
      - part model의 부정확한 score를 통합하는 것을 피하기 위해, 제안된 프레임워크로서 모델 훈련 시 visible parts를 hidden variables로 모델링함.
      - 겹쳐진 parts들에 대한 visible relationship은 수동적으로 정의되거나 독립적으로 가정되는 대신에, discriminative한 deep models을 이용하여 학습시킴.
    - 관련 연구 12(https://www.cv-foundation.org/openaccess/content_cvpr_2013/papers/Ouyang_Single-Pedestrian_Detection_Aided_2013_CVPR_paper.pdf)
      - 혼잡한 보행자들로부터 형성된 unique한 시각 정보를 capture하기 위해 mixture network를 제안함.
      - single pedestrian detectors의 최종 예측을 보강하기 위해, single 및 multiple pedestrian detector들에 의해 추정된 구성(configurations) 사이의 관계를 모델링하도록 확률론적인(probabilistic) 프레임워크가 학습되어짐.
    - 관련 연구 13(OR-CNN, https://arxiv.org/abs/1807.08407)
      - 보행자의 prior 구조 정보를 가시성(visibility) 예측과 함께 최종적인 특징 표현으로 통합한 가려짐을 인지하고 있는(aware) ROI Pooling layer를 제안함.
      - 원래 region을 5개의 part로 나뉘었으며, 각 part에 대해 sub-network는 더 나은 표현을 위해 학습된 가시성 score를 통해 원래의 region feature를 보강시킴.
    - 관련 연구 14(https://cse.buffalo.edu/~jsyuan/papers/2018/Bi-box%20Regression%20for%20Pedestrian%20Detection.pdf)
      - 2개의 bounding boxes들(하나는 전신, 다른 하나는 visible part)을 regression시킴으로써, 보행자 검출과 visible part를 동시에 추정할 수 있는 방법을 제안함.
      - positive-instance 샘플링 기준(criterion)이 visible 영역이 큰 positive 훈련 example들을 bias 시킬 수 있도록 제안되었으며, 가려짐을 인지하고 있는(occlusion-aware) detector를 훈련시키는데 효율적임을 확인함.

3. 기타 Application

- object detection 기술을 사용하는 logo detection, video object detection, vehicle detection, traffic-sign detection, skeleton detection 등의 다양한 실세계 application이 있음.
- **1) logo detection**
  - e-commerce 시스템에서 중요한 연구 주제로서, generic detection과 비교했을 때 logo instance는 비정형 변환(non-rigid transformation)이 심하며 작은 특징이 있음.
  - 관련 연구 1(https://arxiv.org/abs/1803.11417)
    - 잡음이 있는 웹 이미지에서 자동으로 정보를 수집하고 제한된 annotated data를 이용하여 모델을 학습시키는 웹 데이터 학습 원리를 적용함.
  - 관련 연구 2(https://arxiv.org/abs/1612.09322)
    - 제한된 logo instances를 이용하여 detector를 성공적으로 학습시키기 위한 이미지 합성 방법을 제안함.
  - 관련 연구 3(LOGO-Net, https://arxiv.org/abs/1511.02462)
    - e-commerce 웹사이트에서 대규모의 logo 데이터셋을 수집하고 logo detection 시 발생할 수 있는 문제들에 대해서 포괄적인 분석을 수행함.
- **2) video object detection**
  - 존재하는 detection 알고리즘들은 대부분 still 이미지를 위해 고안되었으며, object detection용 video에 직접적으로 적용하는 것은 최선의 선택이 아님.
  - video에서 object를 검출할 때 generic detection과 다른 2개(시간 및 문맥 정보)의 차이점이 있음.
    - 비디오에서 나타나는 object들의 위치 및 외관(appearance)은 인접한 프레임들 간에 시간적으로 일관성을 갖게 됨.
    - 단일 still 이미지와 비교했을 때, 비디오는 수백개의 프레임들로 구성되어 있으므로 훨씬 더 많은 문맥 정보를 포함하고 있음.
  - 관련 연구 1(Seq-NMS, https://arxiv.org/abs/1602.08465)
    - still 이미지의 검출 결과를 sequence들과 연관시킴(associate).
    - 같은 sequence의 box들은 프레임 전체에 걸친 평균 score로 re-scoring되며, sequence를 따르는 다른 boxes들은 NMS에 의해서 억제되어짐(suppressed).
  - 관련 연구 2(T-CNN, https://arxiv.org/abs/1604.04053)
    - Faster R-CNN으로부터 확장된 방법이며, tubelets(시간 경과에 따른 box들의 sequence)으로부터 시간 및 문맥 정보를 통합시킴.
    - optical flow를 이용하여 detection 결과를 인접한 프레임에 전파(propagate)시켰으며, high confidence를 갖는 bounding boxes들로부터 tracking 알고리즘을 적용하여 tubelets들을 생성함.
    - tubelets들을 따르는 boxes들은 tubelets 분류에 기반하여 re-scoring됨.
- **3) 기타 연구들**
  - vehicle detection(DAVE(https://arxiv.org/abs/1607.04564), https://arxiv.org/abs/1709.02480, ShuffleDet(https://arxiv.org/abs/1811.06318))
  - traffic-sign detection(https://zpascal.net/cvpr2016/Zhu_Traffic-Sign_Detection_and_CVPR_2016_paper.pdf, https://arxiv.org/abs/1806.07987)
  - skeleton detection(SRN(https://arxiv.org/abs/1703.02243), https://arxiv.org/abs/1603.09446)

Detection Benchmarks

1. Generic Object Detection Benchmarks

- 아래 그림은 Pascal VOC, MSCOCO, Open Images, LVIS 등의 데이터셋에 대한 일부 example들을 보여주고 있음.

- Pascal VOC2007(http://host.robots.ox.ac.uk/pascal/VOC/pubs/everingham10.pdf)
  - object detection을 위한 중간 규모의 데이터셋으로서 20개의 카테코리를 가지고 있음.
  - training(2501개), validation(2510개), test(5011개)의 이미지들이 있음.

- Pascal VOC2012(http://host.robots.ox.ac.uk/pascal/VOC/voc2012/)
  - object detection을 위한 중간 규모의 데이터셋으로서 Pascal V0C2007과 동일한 20개의 카테코리를 가지고 있음.
  - training(5717개), validation(5823개), test(10991개)의 이미지들이 있음.
  - test set에 대한 annotation 정보는 제공하지 않음.

- MSCOCO(https://arxiv.org/pdf/1405.0312.pdf)
  - 대규모의 데이터셋으로서 80개의 카테고리를 가지고 있음.
  - training(118287개), validation(5000개), test(40670개)의 이미지들이 있음.
  - test set에 대한 annotation 정보는 제공하지 않음.

- Open Images(https://arxiv.org/abs/1811.00982)
  - 198,000개의 이미지가 있으며, 600개의 카테고리에 대해 15,400,000개의 object들을 가지고 있음.
  - detection 벤치마크를 평가하는데 가장 빈도가 높은 500개의 카테고리들이 사용되며, 이러한 카테고리들의 70% 이상이 1,000개 이상의 훈련 샘플을 가지고 있음.

- LVIS(http://openaccess.thecvf.com/content_CVPR_2019/papers/Gupta_LVIS_A_Dataset_for_Large_Vocabulary_Instance_Segmentation_CVPR_2019_paper.pdf)
  - 164,000개의 이미지가 있으며, 1000개 이상의 카테고리를 가지고 있음.
  - 총 2,200,000개의 고품질의 instance segmentation mask를 가지고 있음.
  - 향 후 challenging한 detection, segmentation, low-shot learning task 등을 위한 벤치마크가 될 수 있음.

- ImageNet(http://vision.stanford.edu/pdf/ImageNet_CVPR2009.pdf)
  - 200개의 카테고리를 가지고 있는 중요한 데이터 셋임.
  - 규모는 상당히 크며, object의 스케일 범위는 VOC 데이터셋과 유사하므로 일반적으로 detection 알고리즘을 위한 벤치마크로서 사용되지 않음.

- Evaluation Metrics
  - 검출 정확도와 추론 속도 모두 detection 알고리즘을 평가를 위해서 사용됨.
  - 검출 정확도를 위해서 모든 challenge들에서는 mAP(mean Average Precision)가 평가 측정 방식(evaluation metric)으로 사용됨.
  - VOC2012, VOC2007, ImageNet에서는 IoU 임계치가 0.5인 mAP로 설정하며, MSCOCO의 경우에는 보다 포괄적인 평가 측정 방식이 적용됨.
  - MSCOCO는 detection 알고리즘의 서로 다른 수용력을 보여주는 6개의 평가 score를 가지고 있으며, 서로 다른 IoU 임계치와 서로 다른 스케일의 object들에 대한 성능을 포함시킴.

- Evaluation Metrics
  - 아래 그림은 generic object detection에서 사용하는 평가 측정 방식들을 나타내고 있음.

Alias	Meaning	Definition and Description
FPS	Frame per second	The number of images processed per second.
Ω	IoU threshold	The IoU threshold to evaluate localization.
D_γ	All Predictions	Top γ predictions returned by the detectors with highest confidence score.
TP_γ	True Positive	Correct predictions from sampled predictions
FP_γ	False Positive	False predictions from sampled predictions.
P_γ	Precision	The fraction of TP_γ out of D_γ.
R_γ	Recall	The fraction of TP_γ out of all positive samples.
AP	Average Precision	Computed over the different levels of recall by varying the γ.
mAP	mean AP	Average score of AP across all classes.
TPR	True Positive Rate	The fraction of positive rate over false positives.
FPPI	FP Per Image	The fraction of false positive for each image.
MR	log-average missing rate	Average miss rate over different FPPI rates evenly spaced in log-space

Generic Object Detection			
mAP	mean Average Precision	VOC2007	mAP at 0.50 IoU threshold over all 20 classes.
		VOC2012	mAP at 0.50 IoU threshold over all 20 classes.
		OpenImages	mAP at 0.50 IoU threshold over 500 most frequent classes.
		MSCOCO	• AP_{coco}: mAP averaged over ten Ω: $\{0.5 : 0.05 : 0.95\}$; • AP_{50}: mAP at 0.50 IoU threshold; • AP_{75}: mAP at 0.75 IoU threshold; • AP_S: AP_{coco} for small objects of area smaller than 32^2; • AP_M: AP_{coco} for objects of area between 32^2 and 96^2; • AP_L: AP_{coco} for large objects of area bigger than 96^2;

2. Face Detection Benchmarks

- face detection을 위해 널리 사용되는 데이터셋으로는 WIDER FACE, FDDB, PASCAL FACE 등이 있음.

- WIDER FACE(http://mmlab.ie.cuhk.edu.hk/projects/WIDERFace/support/paper.pdf)
  - 32,203개의 이미지를 가지고 있으며, 큰 범위의 스케일을 갖는 400,000개의 얼굴들이 있음.
  - training(40%), validation(10%), test(50%)로 나뉨.
  - training과 validation set은 온라인으로 이용 가능하며, detection task의 난이도에 따라서 Easy, Medium, Hard 등의 3개로 나누어져 있음.

- FDDB(https://www.researchgate.net/publication/266298783_FDDB_A_Benchmark_for_Face_Detection_in_Unconstrained_Settings)
  - 2,845개의 이미지를 가지고 있으며, 5,171개의 얼굴들이 있음.
  - 일반적으로 face detector들은 WIDER FACE 등의 대규모 데이터셋을 이용하여 먼저 훈련시킨 후 FDDB를 이용하여 test를 수행함.

- PASCAL FACE(http://host.robots.ox.ac.uk/pascal/VOC/pubs/everingham10.pdf)
  - PASCAL person layout test에서 수집되었으며, 851개의 이미지를 가지고 있으며, 1,335개의 label이 있는 얼굴들이 있음.
  - FDDB와 유사하게 일반적으로 test 용도로 사용됨.

- Evaluation Metrics
  - WIDER FACE와 PASCAL FACE는 IoU 임계치가 0.5인 mAP로 설정하며, WIDER FACE의 경우 각 난이도에 따른 결과를 나타낼 수 있음.
  - FDDB의 경우 true positive rate를 이용하며, 평가를 위한 2가지의 annotation(bounding box 수준, eclipse(가림) 수준)을 가지고 있음.

- Evaluation Metrics
  - 아래 그림은 face detection에서 사용하는 평가 측정 방식들을 나타내고 있음.

Alias	Meaning	Definition and Description
FPS	Frame per second	The number of images processed per second.
Ω	IoU threshold	The IoU threshold to evaluate localization.
D_γ	All Predictions	Top γ predictions returned by the detectors with highest confidence score.
TP_γ	True Positive	Correct predictions from sampled predictions
FP_γ	False Positive	False predictions from sampled predictions.
P_γ	Precision	The fraction of TP_γ out of D_γ.
R_γ	Recall	The fraction of TP_γ out of all positive samples.
AP	Average Precision	Computed over the different levels of recall by varying the γ.
mAP	mean AP	Average score of AP across all classes.
TPR	True Positive Rate	The fraction of positive rate over false positives.
FPPI	FP Per Image	The fraction of false positive for each image.
MR	log-average missing rate	Average miss rate over different FPPI rates evenly spaced in log-space

Face Detection			
mAP	mean Average Precision	Pascal Face	mAP at 0.50 IoU threshold.
		AFW	mAP at 0.50 IoU threshold.
		WIDER FACE	• mAP_{easy}: mAP for easy level faces; • mAP_{mid}: mAP for mid level faces; • mAP_{hard}: mAP for hard level faces;
TPR	True Positive Rate	FDDB	• TPR_{dis} with 1k FP at 0.50 IoU threshold, with bbox level. • TPR_{cont} with 1k FP at 0.50 IoU threshold, with eclipse level.

3. Pedestrian Detection Benchmarks

- pedestrian detection을 위해 널리 사용되는 데이터셋으로는 CityPersons, Caltech, ETH, INRIA, KITTI 등이 있음.

- CityPersons(https://arxiv.org/abs/1702.05693)
  - semantic segmentation 데이터셋인 CityScapes(https://arxiv.org/abs/1604.01685) 위에서 구축된 pedestrian detection용 데이터셋으로서, 독일의 여러 도시에서 5000개의 이미지를 capture함.
  - 13,000개의 추가적인 ignored regions를 가지고 있는 총 35,000 명의 사람이 있으며, 모든 사람의 bounding boxes에 대한 annotation과 visible part에 대한 annotation이 함께 제공됨.

- Caltech(http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.402.6884&rep=rep1&type=pdf)
  - pedestrian detection을 위한 가장 인기 있고 도전적인 데이터셋임.
  - LA 지역을 주행하며 30Hz의 VGA 해상도로 10시간 동안 녹화되었으며 training set은 42,782개의 프레임으로, test set은 4,024개의 프레임으로 구성되어 있음.

- ETH(http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.212.8331&rep=rep1&type=pdf)
  - 3개의 비디오 클립이 있으며, 1,804개의 프레임들을 포함하고 있음.
  - 일반적으로 CityPersons과 같은 대규모의 데이터셋으로 훈련을 시킨 모델의 성능 평가를 위한 test set으로서 사용됨.

- INRIA(https://lear.inrialpes.fr/people/triggs/pubs/Dalal-cvpr05.pdf)
  - 대부분 휴가 사진들로부터 수집된 고해상도의 보행자 이미지를 포함하고 있으며, 1,832개의 training set과 288개의 test set이 있음.
  - 특히, training set은 614개의 positive 이미지와 1,218개의 negative 이미지들로 구성되어 있음.

- KITTI(http://ww.cvlibs.net/publications/Geiger2013IJRR.pdf)
  - 1242 x 375 해상도를 갖는 7,481개의 label이 있는 training set 이미지들과 7,518개의 test set 이미지들이 있음.
  - person class는 보행자와 자전거를 타는 사람 등의 2개의 하위 class로 나뉘게 되며, 모두 mAP를 이용하여 평가가 됨.
  - easy, moderate, hard 등의 3개의 평가 측정 방식을 포함하고 있음.

- Evaluation Metrics
  - CityPersons, INRIA, ETH 등은 1e-2에서 100 FPPI(False Positive Per Image) 범위를 갖는 9 point 이상의 log-average miss rate를 이용하여 detector의 성능 평가를 함(낮을수록 좋음).
  - KITTI의 경우 IoU 임계치가 0.5인 표준 mAP가 사용됨.

- Evaluation Metrics
  - 아래 그림은 pedestrian detection에서 사용하는 평가 측정 방식들을 나타내고 있음.

Alias	Meaning	Definition and Description
FPS	Frame per second	The number of images processed per second.
Ω	IoU threshold	The IoU threshold to evaluate localization.
D_γ	All Predictions	Top γ predictions returned by the detectors with highest confidence score.
TP_γ	True Positive	Correct predictions from sampled predictions
FP_γ	False Positive	False predictions from sampled predictions.
P_γ	Precision	The fraction of TP_γ out of D_γ.
R_γ	Recall	The fraction of TP_γ out of all positive samples.
AP	Average Precision	Computed over the different levels of recall by varying the γ.
mAP	mean AP	Average score of AP across all classes.
TPR	True Positive Rate	The fraction of positive rate over false positives.
FPPI	FP Per Image	The fraction of false positive for each image.
MR	log-average missing rate	Average miss rate over different FPPI rates evenly spaced in log-space

Pedestrian Detection			
mAP	mean Average Precision	KITTI	• mAP_{easy}: mAP for easy level pedestrians; • mAP_{mid}: mAP for mid level pedestrians; • mAP_{hard}: mAP for hard level pedestrians;
MR	log-average miss rate	CityPersons	MR: ranging from $1e^{-2}$ to 100 FPPI
		Caltech	MR: ranging from $1e^{-2}$ to 100 FPPI
		ETH	MR: ranging from $1e^{-2}$ to 100 FPPI
		INRIA	MR: ranging from $1e^{-2}$ to 100 FPPI

SOTA for Generic Object Detection

1. Pascal VOC2007 & Pascal VOC2012 결과

- Pascal VOC2007 및 Pascal VOC2012 데이터셋을 이용한 성능 평가 결과
  - 이미지 당 2개 또는 3개의 object들이 있으며, 중간 규모의 데이터셋으로서 object의 크기에 대한 범위는 크지 않음.
  - VOC2007 결과: VOC2007과 VOC2012 trainval set을 이용하여 training 되었으며, VOC2007 test set을 이용하여 test됨.
  - VOC2012 결과: VOC2007과 VOC2012 trainval set과 VOC2007 test set을 이용하여 training 되었으며, VOC2012 test set을 이용하여 test됨.

Method	Backbone	Proposed Year	Input size(Test)	mAP (%)	
				VOC2007	VOC2012
Two-stage Detectors:					
R-CNN [2]	VGG-16	2014	Arbitrary	66.0*	62.4†
SPP-net [2]	VGG-16	2014	~ 600 × 1000	63.1*	-
Fast R-CNN [38]	VGG-16	2015	~ 600 × 1000	70.0	68.4
Faster R-CNN [34]	VGG-16	2015	~ 600 × 1000	73.2	70.4
MR-CNN [131]	VGG-16	2015	Multi-Scale	78.2	73.9
Faster R-CNN [1]	ResNet-101	2016	~ 600 × 1000	76.4	73.8
R-FCN [52]	ResNet-101	2016	~ 600 × 1000	80.5	77.6
OHEM [148]	VGG-16	2016	~ 600 × 1000	74.6	71.9
HyperNet [50]	VGG-16	2016	~ 600 × 1000	76.3	71.4
ION [51]	VGG-16	2016	~ 600 × 1000	79.2	76.4
CRAFT [153]	VGG-16	2016	~ 600 × 1000	75.7	71.3†
LocNet [149]	VGG-16	2016	~ 600 × 1000	78.4	74.8†
R-FCN w DCN [97]	ResNet-101	2017	~ 600 × 1000	82.6	-
CoupleNet [125]	ResNet-101	2017	~ 600 × 1000	82.7	80.4
DeNet512(wide) [94]	ResNet-101	2017	~ 512 × 512	77.1	73.9
FPN-Reconfig [115]	ResNet-101	2018	~ 600 × 1000	82.4	81.1
DeepRegionLet [140]	ResNet-101	2018	~ 600 × 1000	83.3	**81.3**
DCN+R-CNN [132]	ResNet-101+ResNet-152	2018	Arbitrary	**84.0**	81.2
One-stage Detectors:					
YOLOv1 [40]	VGG16	2016	448 × 448	66.4	57.9
SSD512 [42]	VGG-16	2016	512 × 512	79.8	78.5
YOLOv2 [41]	Darknet	2017	544 × 544	78.6	73.5
DSSD513 [112]	ResNet-101	2017	513 × 513	81.5	80.0
DSOD300 [107]	DS/64-192-48-1	2017	300 × 300	77.7	76.3
RON384 [120]	VGG-16	2017	384 × 384	75.4	73.0
STDN513 [111]	DenseNet-169	2018	513 × 513	80.9	-
RefineDet512 [92]	VGG-16	2018	512 × 512	81.8	**80.1**
RFBNet512 [108]	VGG16	2018	512 × 512	**82.2**	-
CenterNet [64]	ResNet101	2019	512 × 512	78.7	-
CenterNet [64]	DLA [64]	2019	512 × 512	80.7	-

* This entry reports the the model is trained with VOC2007 `trainval` sets only.
† This entry reports the the model are trained with VOC2012 `trainval` sets only .

- **Two-stage Detectors**
  - R-CNN(https://arxiv.org/abs/1311.2524)
  - SPP-net(https://arxiv.org/abs/1406.4729)
  - Fast R-CNN(https://arxiv.org/abs/1504.08083)
  - Faster R-CNN(https://arxiv.org/abs/1506.01497)
  - MR-CNN(https://arxiv.org/pdf/1505.01749.pdf)
  - Faster R-CNN(https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf)
  - R-FCN(https://arxiv.org/abs/1605.06409)
  - OHEM(https://arxiv.org/abs/1604.03540)
  - HyperNet(https://arxiv.org/abs/1604.00600)
  - ION(https://arxiv.org/abs/1512.04143)
  - CRAFT(https://arxiv.org/abs/1604.03239)
  - LocNet(https://arxiv.org/abs/1511.07763)
  - R-FCN w DCN(https://arxiv.org/abs/1703.06211)
  - CoupleNet(https://arxiv.org/abs/1708.02863)
  - DeNet512(wide)(https://arxiv.org/abs/1703.10295)
  - FPN-Reconfig(https://arxiv.org/abs/1808.07993)
  - DeepRegionLet(https://arxiv.org/abs/1712.02408)
  - DCN+R-CNN(https://arxiv.org/abs/1803.06799)
- **One-stage Detectors**
  - YOLOv1(https://pjreddie.com/media/files/papers/yolo.pdf)
  - SSD512(https://arxiv.org/abs/1512.02325)
  - YOLOv2(https://arxiv.org/abs/1612.08242)
  - DSSD513(https://arxiv.org/abs/1701.06659)
  - DSOD300(https://arxiv.org/abs/1708.01241)
  - RON384(https://arxiv.org/abs/1707.01691)
  - STDN513(http://openaccess.thecvf.com/content_cvpr_2018/CameraReady/1376.pdf)
  - RefineDet512(https://arxiv.org/abs/1711.06897)
  - RFBNet512(https://arxiv.org/abs/1711.07767)
  - CenterNet(https://arxiv.org/abs/1904.07850)
  - CenterNet(https://arxiv.org/abs/1904.07850)

2. MSCOCO 결과

- MSCOCO 데이터셋
  - 이미지 당 거의 10개 정도의 object들이 있으며, 대다수의 object들은 커다란 크기 변화를 갖는 작은 object들로 구성되어 있음.
  - 따라서, detection 알고리즘의 성능 평가를 위한 challenging한 데이터셋임.
  - "++"은 다중 스케일 테스트, horizontal flip 등의 추론(inference) 전략이 적용되었음을 의미

Method	Backbone	Year	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-Stage Detectors:								
Fast R-CNN [38]	VGG-16	2015	19.7	35.9	-	-	-	-
Faster R-CNN [34]	VGG-16	2015	21.9	42.7	-	-	-	-
OHEM [148]	VGG-16	2016	22.6	42.5	22.2	5.0	23.7	37.9
ION [51]	VGG-16	2016	23.6	43.2	23.6	6.4	24.1	38.3
OHEM++ [148]	VGG-16	2016	25.5	45.9	26.1	7.4	27.7	40.3
R-FCN [52]	ResNet-101	2016	29.9	51.9	-	10.8	32.8	45.0
Faster R-CNN+++ [1]	ResNet-101	2016	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [39]	ResNet-101	2016	36.2	59.1	39.0	18.2	39.0	48.2
DeNet-101(wide) [94]	ResNet-101	2017	33.8	53.4	36.1	12.3	36.1	50.8
CoupleNet [125]	ResNet-101	2017	34.4	54.8	37.2	13.4	38.1	50.8
Faster R-CNN by G-RMI [167]	Inception-ResNet-v2	2017	34.7	55.5	36.7	13.5	38.1	52.0
Deformable R-FCN [52]	Aligned-Inception-ResNet	2017	37.5	58.0	40.8	19.4	40.1	52.5
Mask-RCNN [3]	ResNeXt-101	2017	39.8	62.3	43.4	22.1	43.2	51.2
umd_det [236]	ResNet-101	2017	40.8	62.4	44.9	23.0	43.4	53.2
Fitness-NMS [152]	ResNet-101	2017	41.8	60.9	44.9	21.5	45.0	57.5
DCN w Relation Net [138]	ResNet-101	2018	39.0	58.6	42.9	-	-	-
DeepRegionlets [140]	ResNet-101	2018	39.3	59.8	-	21.7	43.7	50.9
C-Mask RCNN [141]	ResNet-101	2018	42.0	62.9	46.4	23.4	44.7	53.8
Group Norm [237]	ResNet-101	2018	42.3	62.8	46.2	-	-	-
DCN+R-CNN [132]	ResNet-101+ResNet-152	2018	42.6	65.3	46.5	26.4	46.1	56.4
Cascade R-CNN [49]	ResNet-101	2018	42.8	62.1	46.3	23.7	45.5	55.2
SNIP++ [98]	DPN-98	2018	45.7	67.3	51.1	29.3	48.8	57.1
SNIPER++ [146]	ResNet-101	2018	46.1	67.0	51.6	29.6	48.9	58.1
PANet++ [238]	ResNeXt-101	2018	47.4	67.2	51.8	30.1	51.7	60.0
Grid R-CNN [151]	ResNeXt-101	2019	43.2	63.0	46.6	25.1	46.5	55.2
DCN-v2 [144]	ResNet-101	2019	44.8	66.3	48.8	24.4	48.1	59.6
DCN-v2++ [144]	ResNet-101	2019	46.0	67.9	50.8	27.8	49.1	59.5
TridentNet [239]	ResNet-101	2019	42.7	63.6	46.5	23.9	46.6	56.6
TridentNet [239]	ResNet-101-Deformable	2019	**48.4**	**69.7**	**53.5**	31.8	51.3	60.3

- **Two-stage Detectors**
  - Fast R-CNN(https://arxiv.org/abs/1504.08083)
  - Faster R-CNN(https://arxiv.org/abs/1506.01497)
  - OHEM(https://arxiv.org/abs/1604.03540)
  - ION(https://arxiv.org/abs/1512.04143)
  - OHEM++(https://arxiv.org/abs/1604.03540)
  - R-FCN(https://arxiv.org/abs/1605.06409)
  - Faster R-CNN+++(https://arxiv.org/abs/1512.03385)
  - Faster R-CNN w FPN(https://arxiv.org/abs/1612.03144)
  - DeNet-101(wide)(https://arxiv.org/abs/1703.10295)
  - CoupleNet(https://arxiv.org/abs/1708.02863)
  - Faster R-CNN by G-RMI(https://arxiv.org/abs/1611.10012)
  - Deformable R-FCN(https://arxiv.org/abs/1703.06211)
  - Mask-RCNN(https://arxiv.org/abs/1703.06870)
  - umd_det(https://arxiv.org/abs/1704.04503)
  - Fitness-NMS(https://arxiv.org/abs/1711.00164)
  - DCN w Relation Net(https://arxiv.org/abs/1711.11575)
  - DeepRegionlet(https://arxiv.org/abs/1712.02408)
  - C-Mask RCNN(http://openaccess.thecvf.com/content_ECCV_2018/papers/Zhe_Chen_Context_Refinement_for_ECCV_2018_paper.pdf)
  - Group Norm(https://eccv2018.org/openaccess/content_ECCV_2018/papers/Yuxin_Wu_Group_Normalization_ECCV_2018_paper.pdf)
  - DCN+R-CNN(https://arxiv.org/abs/1803.06799)
  - Cascade R-CNN(https://arxiv.org/abs/1712.00726)
  - SNIP++(http://openaccess.thecvf.com/content_cvpr_2018/papers/Singh_An_Analysis_of_CVPR_2018_paper.pdf)
  - SNIPER++(https://arxiv.org/abs/1805.09300)
  - PANet++(https://arxiv.org/abs/1803.01534)
  - Grid R-CNN(https://arxiv.org/abs/1811.12030)
  - DCN-v2(https://arxiv.org/abs/1811.11168)
  - DCN-v2++(https://arxiv.org/abs/1811.11168)
  - TridentNet(https://arxiv.org/abs/1901.01892)
  - TridentNet(https://arxiv.org/abs/1901.01892)

● MSCOCO 데이터셋

Single-Stage Detectors:								
SSD512 [42]	VGG-16	2016	28.8	48.5	30.3	10.9	31.8	43.5
RON384++ [120]	VGG-16	2017	27.4	49.5	27.1	-	-	-
YOLOv2 [41]	DarkNet-19	2017	21.6	44.0	19.2	5.0	22.4	35.5
SSD513 [112]	ResNet-101	2017	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513 [112]	ResNet-101	2017	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet800++ [43]	ResNet-101	2017	39.1	59.1	42.3	21.8	42.7	50.2
STDN513 [111]	DenseNet-169	2018	31.8	51.0	33.6	14.4	36.1	43.4
FPN-Reconfig [115]	ResNet-101	2018	34.6	54.3	37.3	-	-	-
RefineDet512 [92]	ResNet-101	2018	36.4	57.5	39.5	16.6	39.9	51.4
RefineDet512++ [92]	ResNet-101	2018	41.8	62.9	45.7	25.6	45.1	54.1
GHM SSD [147]	ResNeXt-101	2018	41.6	62.8	44.2	22.3	45.1	55.3
CornerNet511 [63]	Hourglass-104	2018	40.5	56.5	43.1	19.4	42.7	53.9
CornerNet511++ [63]	Hourglass-104	2018	42.1	57.8	45.3	20.8	44.8	56.7
M2Det800 [116]	VGG-16	2019	41.0	59.7	45.0	22.1	46.5	53.8
M2Det800++ [116]	VGG-16	2019	44.2	64.6	49.3	29.2	47.9	55.1
ExtremeNet [240]	Hourglass-104	2019	40.2	55.5	43.2	20.4	43.2	53.1
CenterNet-HG [64]	Hourglass-104	2019	42.1	61.1	45.9	24.1	45.5	52.8
FCOS [241]	ResNeXt-101	2019	42.1	62.1	45.2	25.6	44.9	52.0
FSAF [95]	ResNeXt-101	2019	42.9	63.8	46.3	26.6	46.2	52.7
CenterNet511 [65]	Hourglass-104	2019	44.9	62.4	48.1	25.6	47.4	57.4
CenterNet511++ [65]	Hourglass-104	2019	**47.0**	**64.5**	**50.7**	28.9	49.9	58.9

- **One-stage Detectors**
  - SSD512(https://arxiv.org/abs/1512.02325)
  - RON384++(https://arxiv.org/abs/1707.01691)
  - YOLOv2(https://arxiv.org/abs/1612.08242)
  - SSD513(https://arxiv.org/abs/1701.06659)
  - DSSD513(https://arxiv.org/abs/1701.06659)
  - RetinaNet800++(https://arxiv.org/abs/1708.02002)
  - STDN513(http://openaccess.thecvf.com/content_cvpr_2018/CameraReady/1376.pdf)
  - FPN-Reconfig(https://arxiv.org/abs/1808.07993)
  - RefineDet512(https://arxiv.org/abs/1711.06897)
  - RefineDet512++(https://arxiv.org/abs/1711.06897)
  - GHM SSD(https://arxiv.org/abs/1811.05181)
  - CornerNet511(https://arxiv.org/abs/1808.01244)
  - CornerNet511++(https://arxiv.org/abs/1808.01244)
  - M2Det800(https://arxiv.org/abs/1811.04533)
  - M2Det800++(https://arxiv.org/abs/1811.04533)
  - ExtremeNet(https://arxiv.org/abs/1901.08043)
  - CenterNet-HG(https://arxiv.org/abs/1904.07850)
  - FCOS(https://arxiv.org/abs/1904.01355)
  - FSAF(https://arxiv.org/abs/1903.00621)
  - CenterNet511(https://arxiv.org/abs/1904.08189)
  - CenterNet511++(https://arxiv.org/abs/1904.08189)

Concluding Remarks and Future Directions

도전적인 연구 주제들

- object detection은 상당히 활발하게 연구되고 있으며, 거의 매달 SOTA 결과들이 보고 되고 있음.
- 하지만 아래와 같이 여전히 open된 도전적인 연구들이 있음.
- **i) 규모에 가변적인(scalable) proposal generation 전략**
  - 현재 대부분의 detector들은 anchor에 기반한 방법들이며, 검출 성능을 제한시키는 결정적인 단점들이 존재함.
  - anchor priors가 수동적으로 설계되었기 때문에 다중 스케일의 objects들을 matching시키기 어려움을 겪게 되며, IoU에 기반한 matching 전략 또한 휴리스틱한 방법임.
  - anchor에 기반한 방법들을 anchor-free한 방법들(예: keypoint 기반 방법들)로 변환시키려는 연구들이 제안되었지만, 높은 계산 비용 등의 커다란 공백을 가진 한계들이 여전히 존재함.
  - anchor-free한 방법의 개발은 object detection에서 상당히 hot한 주제들임.
    - 관련 연구들(CornerNet(https://arxiv.org/abs/1808.01244), FSAF(https://arxiv.org/abs/1903.00621), https://arxiv.org/pdf/1901.08043.pdf, FCOS(https://arxiv.org/abs/1904.01355), CenterNet(https://arxiv.org/abs/1904.08189))
  - 위와 같은 요소를 고려해봤을 때, 효율적이고 효과적인 proposal generation 전략을 설계하는 것은 향 후 연구에 있어서 매우 중요한 방향이 될 것임.

- **ii) 문맥 정보(contextual information)의 효율적인 부호화(encoding)**
  - 시각 세계의 objects들은 강한 관계성(relationships)을 가지고 있으며 문맥은 시각 세계를 좀 더 이해할 수 있는 중요한 역할을 하기 때문에, 문맥은 object detection 결과에 기여하거나 방해를 할 수 있음.
  - 하지만 문맥 정보를 올바르게 사용하는 방법에 초점을 둔 연구들은 거의 없음.
  - 따라서 object detection을 위해 어떻게 문맥 정보를 효율적으로 통합할 수 있는가에 대한 연구들은 향 후에 유망한 연구 주제가 될 것임.

- **iii) 자동 기계 학습(AutoML)에 기반한 detection**
  - 특정한 task를 위해 최적의 backbone architecture를 설계하는 것은 성능을 개선시키는데 도움이 되지만, 상당한 엔지니어링적인 노력이 필요함.
  - 따라서, 데이터셋에 대해 backbone architecture를 직접적으로 학습하는 것은 매우 흥미롭고 중요한 연구 방향일 수 있음.
    - AutoML을 image classification에 적용한 연구
      - https://arxiv.org/abs/1707.07012, EfficientNet(https://arxiv.org/abs/1905.11946)
    - AutoML을 object detection에 적용한 연구들
      - DetNAS(https://arxiv.org/abs/1903.10979v1), NAS-FPN(https://arxiv.org/abs/1904.07392), Data Augmentation Learning(https://arxiv.org/abs/1906.11172)
      - baseline 대비 상당한 성능 향상이 있었지만, 계산에 필요한 자원들은 대부분의 연구자들이 이용할 수 없음(single 모델 훈련을 위해 100개가 넘는 GPU가 필요함).
      - 따라서, 낮은 계산량의 프레임워크 개발은 object detection에 큰 영향을 줄 수 있을 것임.
      - 또한, 향 후에는 detection task에 대한 proposal generation, 영역 부호화 등의 새로운 구조 정책(structure policy)이 활용될 수 있음.

- **iv) object detection용 benchmarks의 발전**
  - 현재 MSCOCO는 detection benchmark를 테스트하기 위해 가장 많이 사용되고 있음.
  - 하지만, MSCOCO는 80개의 카테고리만 있기 때문에 실세계의 복잡한 환경을 이해하기엔 너무나 부족한 편임.
  - LVIS 데이터셋(http://openaccess.thecvf.com/content_CVPR_2019/papers/Gupta_LVIS_A_Dataset_for_Large_Vocabulary_Instance_Segmentation_CVPR_2019_paper.pdf)
    - 보다 많은 카테고리에 대한 정보를 수집하기 위해서 제안됨.
    - 164,000개의 이미지가 있으며, 1000개 이상의 카테고리를 가지고 있음.
    - 총 2,200,000개의 고품질의 instance segmentation mask를 가지고 있음.
    - 많은 카테고리를 가지고 있지만, 카테고리 별 데이터가 부족한 실세계의 low-shot 시나리오를 모방하고 있음.
    - 향 후 challenging한 detection, segmentation, low-shot learning task 등을 위한 benchmark가 될 수 있음.

- **v) low-shot object detection**
  - 제한된 label을 이용하여 detector를 훈련시키는 것을 low-shot detection이라 함.
  - 딥러닝 기반의 detector들은 상당한 양의 파라미터를 가지기 때문에 데이터 부족을 겪게 되며, 만족스러운 성능을 위해서는 상당한 양의 labeling된 데이터를 필요로 함.
  - 하지만, bounding box 수준의 annotation을 이용하여 이미지 내 object들을 labeling하는 것은 상당히 시간 소비적임.
  - low shot learning은 classification task를 위해서는 활발하게 연구되었지만, detection task를 위한 연구는 드문편임.
  - 관련 연구 1(MSPLD, https://arxiv.org/abs/1706.08249)
    - 대규모의 label이 없는 데이터셋을 이용 가능한 상황에서 semi-supervised learning을 이용하여 low shot learning 문제를 해결하고자 함.
  - 관련 연구 2(RepMet, https://arxiv.org/abs/1806.04728)
    - feature embedding 공간과 훈련 set 카테고리의 데이터 분포를 함께 학습시키는 DML(Deep Metric Learning) 구조를 채택함.
    - 하지만, 유사한 개념(concept)을 가진 데이터셋(동물)에 대해서만 테스트가 됨.
  - 관련 연구 3(LSTD, https://arxiv.org/abs/1803.01529)
    - knowledge regularization을 이용하여대량의 annotation된 외부 데이터셋으로부터 target set으로 knowledge를 전이하는 전이 학습(transfer learning)에 기반한 low-shot detection을 제안함.
    - 여전히 overfitting을 보임.
  - low-shot detection task는 개선을 위한 많은 여지가 많이 남아 있음.

- **vi) detection task에 적합한 backbone architecture**
  - detection 문제에 있어서, 대규모의 데이터셋을 이용하여 pretrained된 classification 모델의 weight를 적용하는 것은 패러다임이 되었음.
  - 하지만 여전히 classification과 detection task를 사이의 충돌(conflicts)이 존재하며(DetNet, https://arxiv.org/abs/1804.06215), 이렇게 하는 것이 최적의 해결책은 아님.
  - "MSCOCO 데이터셋을 이용한 성능 평가 결과"에서 살펴볼 수 있는 바와 같이 대부분의 최신 detection 알고리즘들은 classification backbone에 기반하고 있으며, 몇몇만 다른 선택(예: HourglassNet에 기반한 CornerNet)을 취하고 있음.
  - 따라서 detection-aware backbone 구조를 개발하는 것은 향 후의 중요한 연구 방향이라고 할 수 있음.

- **vii) 기타의 다른 연구 주제**
  - 위에서 살펴본 내용 이외에도 다른 open된 연구 분야들이 있음.
  - 관련 연구 1(MegDet, https://arxiv.org/abs/1711.07240)
    - large batch learning과 관련된 연구로서, batch 크기는 DCNN 훈련 시 중요한 요소이지만 detection task에서는 활발하게 연구되지 않았음.
  - 관련 연구 2(https://arxiv.org/abs/1708.06977)
    - incremental learning과 관련된 연구로서, incremental learning은 초기 학습 데이터 없이 새로운 task에 적응시킬 경우 detection 알고리즘은 여전히 치명적인 망각(catastrophic forgetting)을 겪게 됨.
  - 이렇게 open되고 근본적인 연구 문제들은 향 후 연구를 위해 보다 많은 관심을 기울일 필요가 있음.