Deep Learning for Generic Object Detection: A Survey

key factors (1)~(6) which have emerged in

generic object detection based on deep learning

(1) Detection Frameworks : 2 Stage vs 1 Stage :

Two-Stage detectors :

1 특징

높은 탐지 정확도, 구조가 flexible, region based classification에 적합

2 Framworks

2.1 RCNN : AlexNet with Selective Search

Disadvantages : Training is Expensive in space and time

Testing Slow

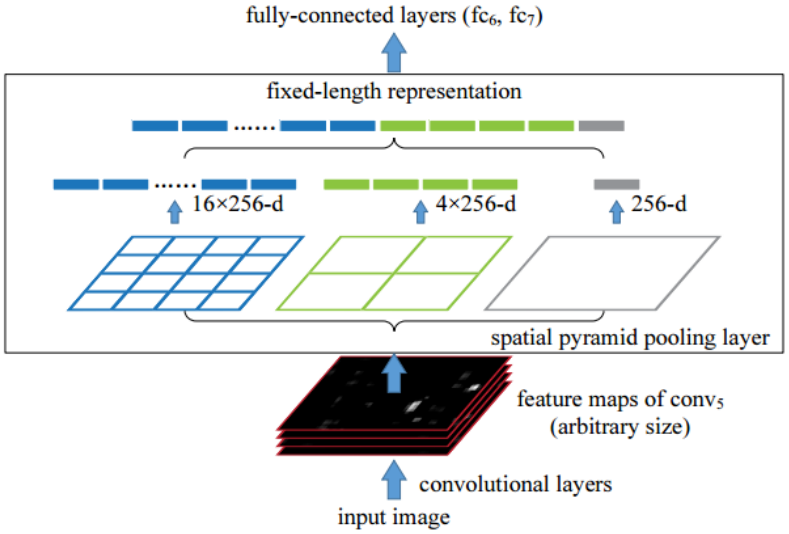
2.2 SPPNet :

FC layer가 Fixed Inputs Size를 필요로 함

Fixed Input Size가 이미지 정보 손실과 변형을 만든다

Spatial Pyramid Pooling layer를 FC layer에 추가하여 Arbitary Input Size 가능

Spatial Pyramid Pooling :



Advantages :

CNN을 이미지에 한번만 적용

Convolution Feature sharing

Disadvantages :

Detector training 의 속도증가가 RCNN과 차이가 거의 없음

Convolution Layer update가 불가능 (very deep network 의 정확도를 제한)

2.3 Fast RCNN

2.4 Faster RCNN

2.5 RFCN (Region based Fully Convolutional Network) :

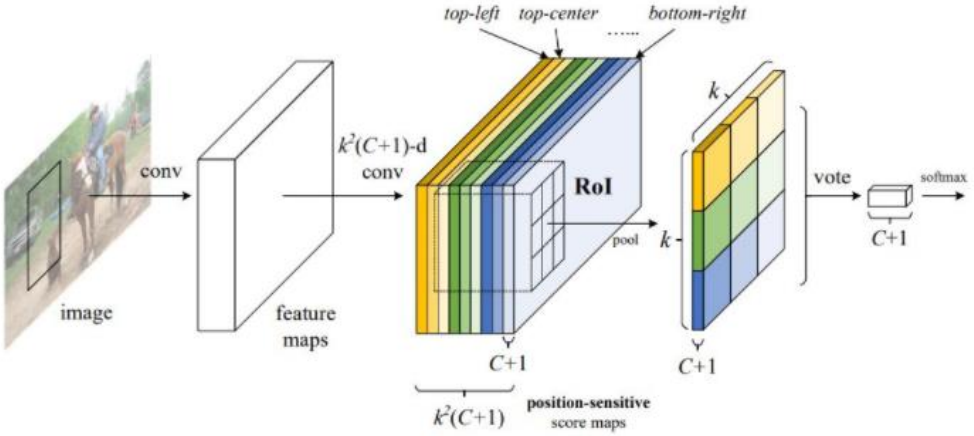
Faster RCNN에서 RoI 별로 Sub network를 통과시켜야 한다

Backbone Network를 통해 얻은 Feature Map은 translation invariance 하다 (위치 정보가 소실된 채로 학습)

FCN을 적용하여 RoI를 구하며 Classification, BBox regression 진행 가능

(Computation Shared over entire image)

Position-sensitive score map -> RoI pooling -> Voting



Disadvantages :

속도 향상이 거의 없음

SPP Layer 이전의 Convolution layers update 불가능

2.6 MaskRCNN

2.7 Chained Cascade Network and Cascade RCNN

IoU가 높으면 많은 positive sample에 사라지기 때문에 overfitting이 발생.

train과 inference시에 작용하는 IoU가 다름. (0.5 IoU로 학습을 한 모델은 COCO metric에 따라 0.5~0.95 IoU 범위로 test 진행).

IoU를 증가시키면서 연속적으로 detector을 학습.

매 stage가 지날 수록 proposal의 정확도는 높아지고, 설정한 IoU도 높아지므로 detector의 성능이 향상

2.8 Light Head RCNN

RFCN의 Detection network의 head를 경량화 하여 RoI computation을 줄임

One-Stage detectors :

1 특징

Two-Stage detectors보다 빠르다(avoid preprocessing algorithm, lightweight backbone),

더 적은 후보 영역으로 예측을 수행하고 분류 하위 네트워크를 완전히 통합.

2 Framworks

2.1 DetectorNet

AlexNet 사용, 마지막 softmax classifier layer을 regression layer로 대체

하나의 네트워크를 사용하여 거친 그리드를 통해 전경 픽셀을 예측

4개의 추가 네트워크를 사용하여 객체의 상단, 하단, 왼쪽 및 오른쪽 절반을 예측

Disadvantages :

DetectorNet은 이미지의 많은 자르기 작업을 수행

모든 자르기 작업에서 각 부분에 대해 여러 네트워크를 실행(속도가 느려짐)

2.2 OverFeat

First One-Stage object detector based on Fully Convolutional deep networks.

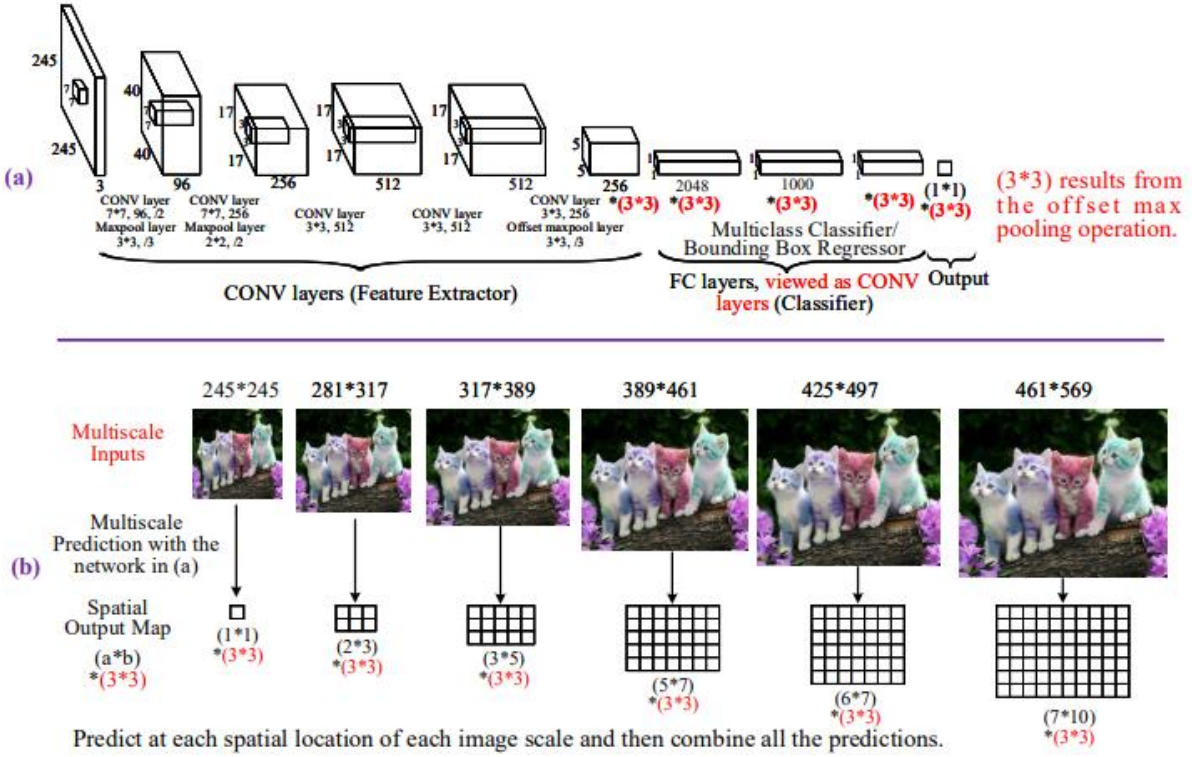
Fixed size input 문제 해결

1. OverFeat는 멀티스케일 기능을 활용, 네트워크를 통해 최대 6개의 확대된 원본 이미지 전달(전체 성능 향상).

2. last Convolution layer 이후 max pooling (해상도 향상)

3. BBox regression : object를 찾게 되면 하나의 BBox regressor 적용하여 classifier와 regressor가 같은 feature extraction layer을 공유(FC layer만 classification network 연산 이후에 한번더 연산)하여 속도가 빠르다.

4. Greedy merge를 통해 각각의 BBox prediction(location, scale) 을 통합

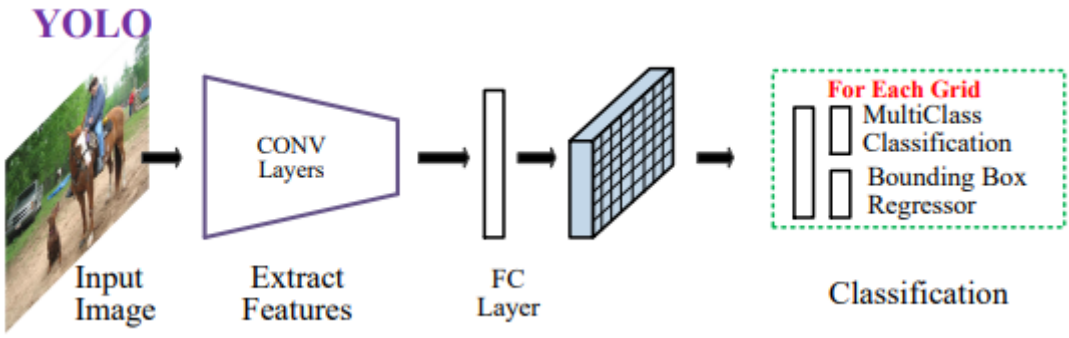


2.3 YOLO (You Only Look Once)

이미지 내의 bounding box와 class probability를 single regression problem으로 간주

Region proposal step 삭제

Single convolutional network를 통해 multiple bounding box에 대한 class probablility를 계산



Input image를 S * S개의 grid로 나누어 각 grid cell 마다

Class probability(C), BBox location (B), confidence scores 를 예측

Advantages :

빠르다 (45 FPS), Fast YOLO (155FPS)

Image 전체를 한 번에 바라보는 방식으로 class에 대한 context 이해도가 높음.

(낮은 background error(False-Positive))

Disadvantages :

Localization Errors than Fast RCNN Especially small objects (grid cell contain one object)

2.4 YOLOv2

GoogleNet 에서 DarkNet19로 변경

Batch normalization 추가

Fully Connected layer 삭제

kmeans 와 multiscale training 을 통한 Good anchor Box 사용

standard detection tasks 에서 SOTA 달성

YOLO9000

9000개 이상의 object categories를 real-time 에 detect

Disadvantages :

Not good at detecting small objects

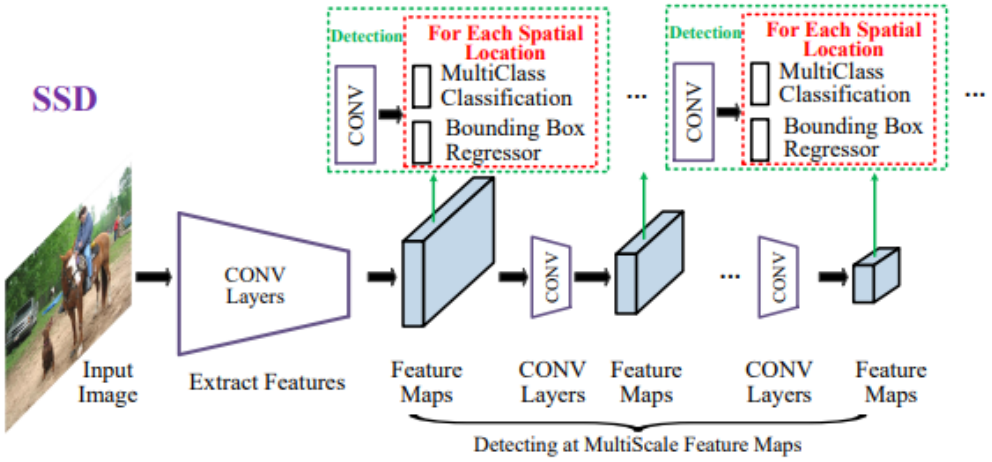
2.5 SSD (Single Shot Multibox Detector)

정확도 향상을 위해 만들어짐

Faster RCNN의 RPN + YOLO + multiscale convolution features

각 Feature Map에서 Object detection 수행(고정 개수의 BBox와 score 예측)

각 수행한 것을 Non Maximum Suppression step을 통해 Final detection



Disadvantages :

Not good at detecting small objects

2.6 CornerNet

Anchor Box의 문제점 (slowing down training, extra hyperparameters)

Detecting paired keypoints (top-left and bottom-right) to BBox prediction

(idea from Associative Embedding in multiperson pose estimation)

Two stacked Hourglass network을 Backbone 으로 사용.

탐지 프레임워크의 설계는 다음과 같은 중요한 설계 선택 항목으로 수렴

- A fully convolutional pipeline
- Exploring complementary information from other correlated tasks, e.g., Mask RCNN [102]
- Sliding windows [229]
- Fusing information from different layers of the backbone.

다단계 객체 감지가 속도 정확성 균형을 위한 미래 프레임워크가 될 수 있다.

(2) Backbone Networks

ResNet, ResNeXt, InceptionResNet : Deeper Backbone(성능 향상 하지만 계산적으로 더 비싸고 교육을 위해 훨씬 더 많은 데이터와 대규모 컴퓨팅이 필요)

MobileNet : 속도 향상에 초점을 맞춤, 계산 비용 및 모델 크기의 1/30만으로 ImageNet에서 VGGNet16 정확도 달성

(3) Improving the Robustness of Object Representation

Deep CNN based detectors 는 위의 table 에 나열된 deep CNN architectures를 backbone network로 사용하며, CNN의 top layer로부터 나온 features를 object representations로 사용한다.

Challenge issue : detecting objects across a large range of scales

Strategy : run the detector over a number of scaled input images

(정확한 탐지, 추론 시간과 memory 제한)

1. Handling of Object Scale Variations

Layers		
High	Large receptive field and strong semantics Robust to variations	Low resolution Loss of geometric details
Low	Small reception field Less sensitive to semantics	High resolution Rich geometric details

Target Object	
Small	requires fine detail information in earlier layers may very well disappear at later layers
Large	semantic concept will emerge in much later layers

여러 CNN 계층을 이용하여 탐지 정확도 향상하기 위해 3가지의 **multiscale object detection**

A. Detecting with combined features of multiple layers

B. Detecting at multiple layers

C. Combinations of the above two methods.

A. Detecting with combined features of multiple layers

Prediction 전에 multiple layers로부터 얻은 feature 을 combine

Concatenates features from different layers

(a). ION

RoI pooling (extract RoI features from multiple layers)

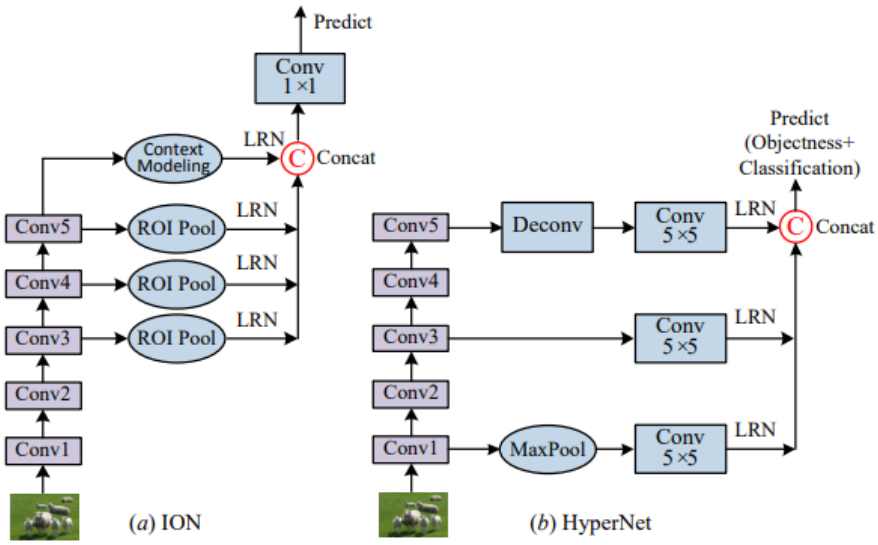
Selective Search (generate object proposals)

Concatenated features (classify edgeboxes)

(b) HyperNet

Integrate deep, intermediate and shallow features (generate object proposals)

End-to-End joint strategy (predict object)



Advantages : Localization and classification

Disadvantages : increased computational complexity

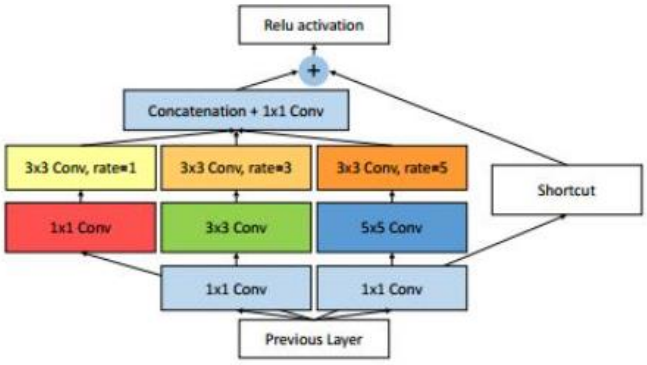
B. Detecting at multiple CNN layers

Different layers의 different resolution에서 object prediction 후 combine

(a) SSD

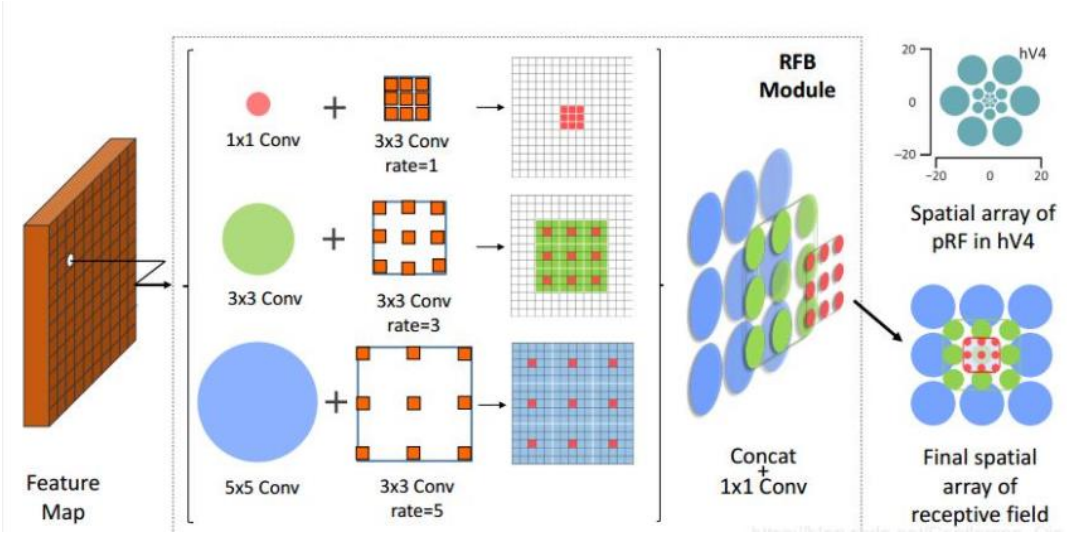
(b) RFBNet

SSD의 마지막 convolution layer를 Receptive Field Block으로 교체



RFB module

Inception module 과 유사하지만 different kernel과 convolution layer로 이루어진 각 branch를 합친다.

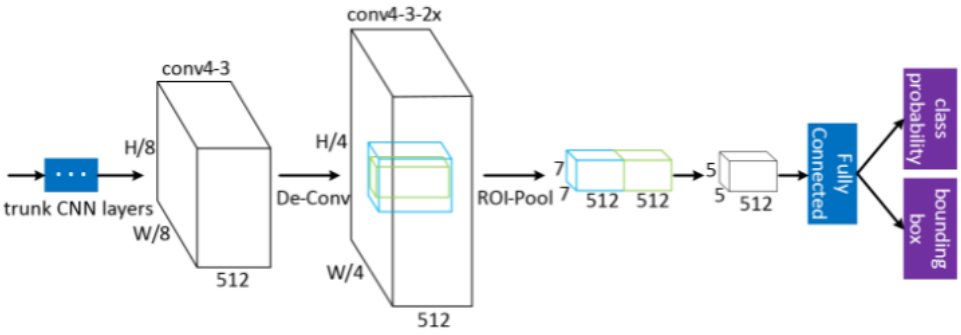


Feature의 discriminability 와 robustness 향상

architecture

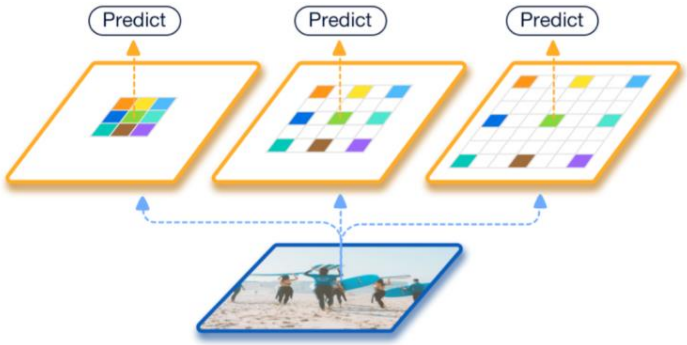
(c) MSCNN

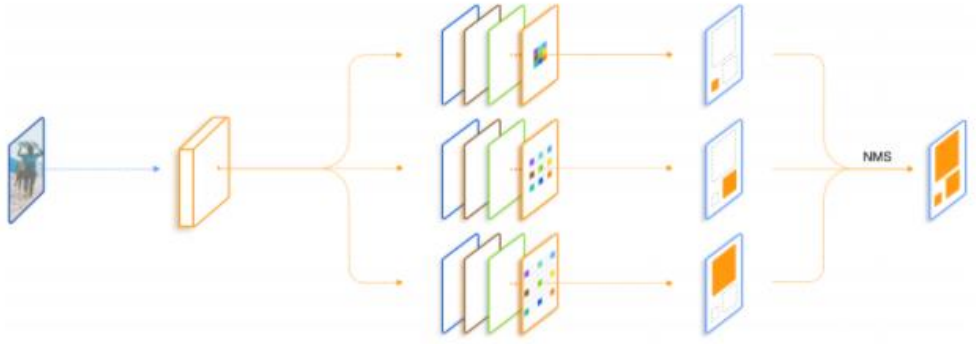
Deconvolution (increase feature map resolution)



(d) TridentNet

Parallel multibranch architecture





같은 transformation parameter를 공유하지만 각자 다른 dilation rate로 convolution 하여 다른 scale의 receptive field을 적용한다. (dilated convolution)

C. Combinations of the above two methods

크기가 거의 같은 feature를 사용하여 서로 다른 scale의 object를 탐지

Downscaled feature map : 큰 객체를 탐지

Upscaled feature map : 작은 객체를 탐지

Multiple layer에서 object detect 후 다른 layers로부터 얻은 features 을 결합하는 것은 segmentation에 매우 효율적

개체 인스턴스 간 스케일 변동 문제를 완화하기 위한 디텍터

(3) Combination of (1) and (2)										
	DSSD [77]	Free	ResNet101	SSD	81.5 (07+12)	80.0 (07T+12)	53.3	33.2	2017	Use Conv-Deconv, as shown in Fig. 17 (c1, c2).
	FPN [167]	RPN	ResNet101	Faster RCNN	–	–	59.1	36.2	CVPR17	Use Conv-Deconv, as shown in Fig. 17 (a1, a2); Widely used in detectors.
	TDM [247]	RPN	ResNet101 VGG16	Faster RCNN	–	–	57.7	36.8	CVPR17	Use Conv-Deconv, as shown in Fig. 17 (b2).
	RON [136]	RPN	VGG16	Faster RCNN	81.3 (07+12+CO)	80.7 (07T+12+CO)	49.5	27.4	CVPR17	Use Conv-deconv, as shown in Fig. 17 (d2); Add the objectness prior to significantly reduce object search space.
	ZIP [156]	RPN	Inceptionv2	Faster RCNN	79.8 (07+12)	–	–	–	IJCV18	Use Conv-Deconv, as shown in Fig. 17 (f1). Propose a map attention decision (MAD) unit for features from different layers.
	STDN [321]	Free	DenseNet169	SSD	80.9 (07+12)	–	51.0	31.8	CVPR18	A new scale transfer module, which resizes features of different scales to the same scale in parallel.
	RefineDet [308]	RPN	VGG16 ResNet101	Faster RCNN	83.8 (07+12)	83.5 (07T+12)	62.9	41.8	CVPR18	Use cascade to obtain better and less anchors. Use Conv-deconv, as shown in Fig. 17 (e2) to improve features.
	PANet [174]	RPN	ResNeXt101 +FPN	Mask RCNN	–	–	**67.2**	**47.4**	CVPR18	Shown in Fig. 17 (g). Based on FPN, add another bottom-up path to pass information between lower and topmost layers; adaptive feature pooling. Ranked 1^{st} and 2^{nd} in COCO 2017 tasks.
	DetNet [164]	RPN	DetNet59+FPN	Faster RCNN	–	–	61.7	40.2	ECCV18	Introduces dilated convolution into the ResNet backbone to maintain high resolution in deeper layers; Shown in Fig. 17 (i).
	FPR [137]	–	VGG16 ResNet101	SSD	82.4 (07+12)	81.1 (07T+12)	54.3	34.6	ECCV18	Fuse task oriented features across different spatial locations and scales, globally and locally; Shown in Fig. 17 (h).
	M2Det [315]	–	SSD	VGG16 ResNet101	–	–	64.6	44.2	AAAI19	Shown in Fig. 17 (j), newly designed top down path to learn a set of multilevel features, recombined to construct a feature pyramid for object detection.

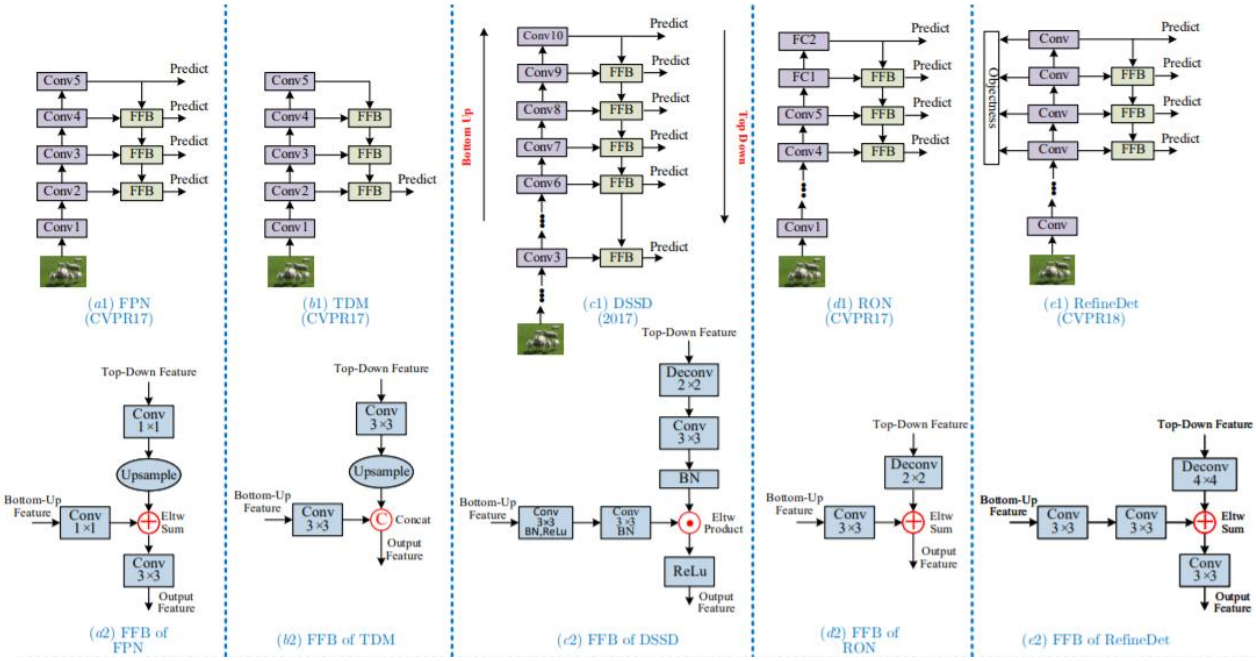
FPN, TDM, DSSD, ZIP, RON, RefineDet

pyramid architecture 사용

(a1-e1) bottom-up, feed-foward network를 보완하기 위해 lateral connection이 있는 하향식 네트워크를 통합하는 매우 유사한 detection architecture를 가짐

특히, bottom-up 통과 후 최종 상위 레벨 semantic feature가 top-down network에 다시 전송되어 lateral processing 이후 intermediate layers에서 생성된 bottom-up feature와결합되고, 결합된 기능이 detection에 사용

(a2-e2) Feature Fusion Block 은 다른 layers로부터 온 features의 선택을 조절하고 multilayer features를 combine



FPN 은 object detection, instance segmentation에 적용되어 큰 성능 발전에 기여

STDN

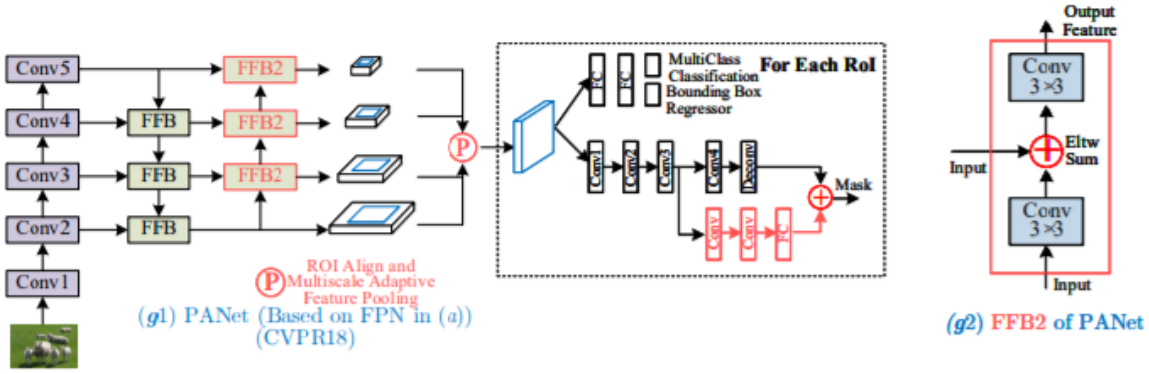
DensNet (different layers의 features combine)

scale transfer module (다른 해상도의 feature map을 가짐)

조금의 추가 cost로 module이 backbone network에 embedded

Improve on the pyramid architectures

PANet

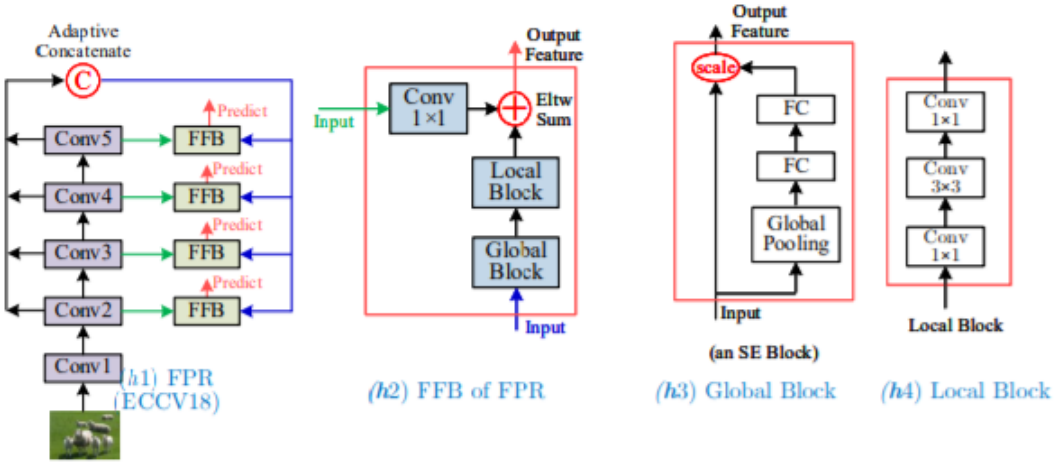


Add another bottom-up part (information path를 줄임, feature pyramid 향상)

Adaptive feature pooling (각 feature level의 proposal 로부터 얻어진 feature 집계)

Complementary branch가 각 proposal의 다른 관점을 취득 (improve mask prediction)

FPR



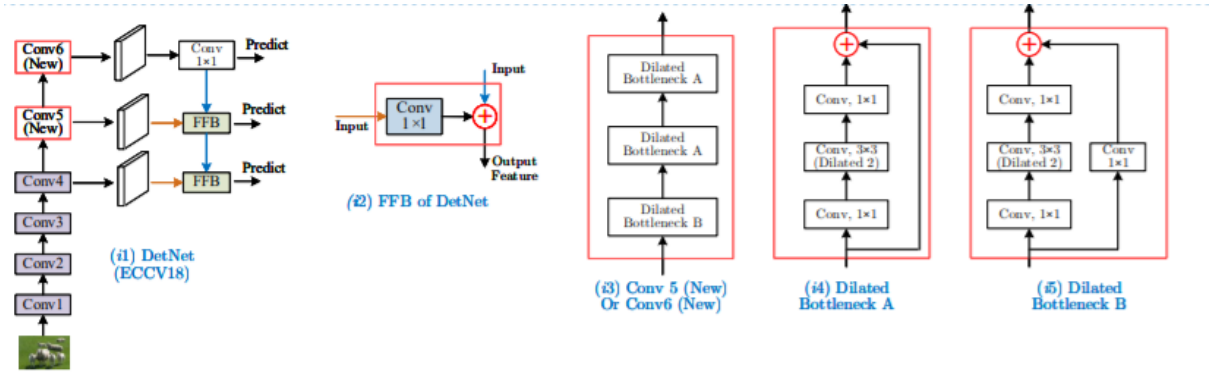
Feature pyramid 구조 변형

Highly Nonlinear 하지만 효율적

Adaptive Concatenate

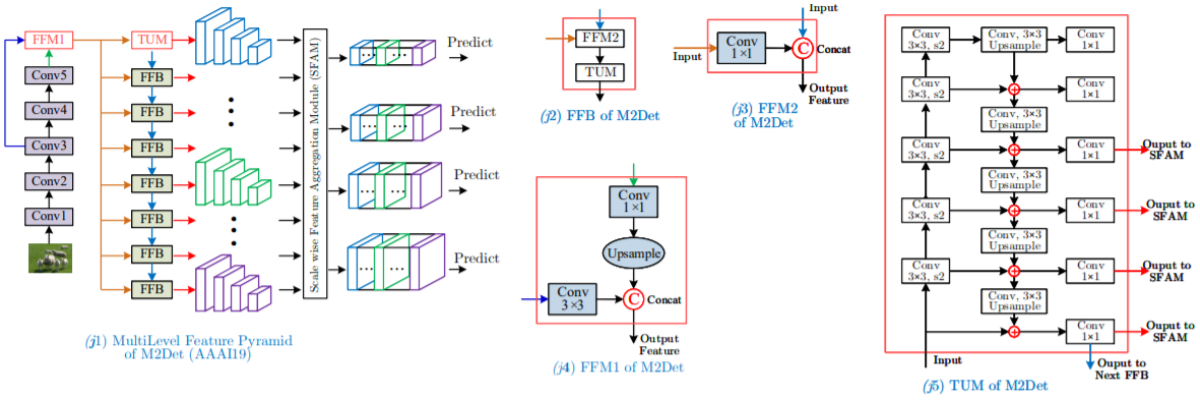
Strong FFB module (모든 scale에 strong semantics를 전달)

DetNet



Dilated Convolution (deep layer에서의 높은 공간 해상도 유지)

M2Det



MultiLevel Feature Pyramid Network를 SSD에 결합

TUM

2. Handling of Other Intraclass Variations

1.에서 다룬 scale 말고도 real-world 변화에 robustness 하기위해 아래의 것이 필요

- **Geometric transformations**

DNNs는 기하학적 변환에 대하여 spatially invariant 하지 못함

Local max pooling layers 통해 translation invariance

중간의 feature maps 은 input data의 큰 기하학적 변환에 invariant 하지 못함

scale, rotation 아니면 다른 유형의 변환에 대해 불변적인 CNN 표현을 학습

(enhance robustness)

Rotation invariance에 대한 연구는 제한적 (대중적인 dataset이 rotated images를 미 제공)

Spatial Transformer Network (STN)

affine parameter를 학습하여 spatially invariant 하게 함

rotated text detection, rotated face detection, generic object detection에 사용

Deformable Part based Models (DPMs)

변형 가능한 구성으로 배열된 부품별로 객체를 나타내는 일반적인 객체 감지

DPM 모델링은 객체 자세, 관점 및 비강체 변형에 대한 변환에 덜 민감

연구자가 객체 구성을 명시적으로 모델링하여,

CNN 기반 탐지를 개선하도록 동기를 부여

sliding window technique 사용

region proposal 없이 AlexNet에서 학습한 심층 feature을 사용하여

DPMs을 CNN과 결합

Deformable Convolution Networks (DCN)

고정된 수용영역에서 특징을 추출하면 translation-invariance 생김

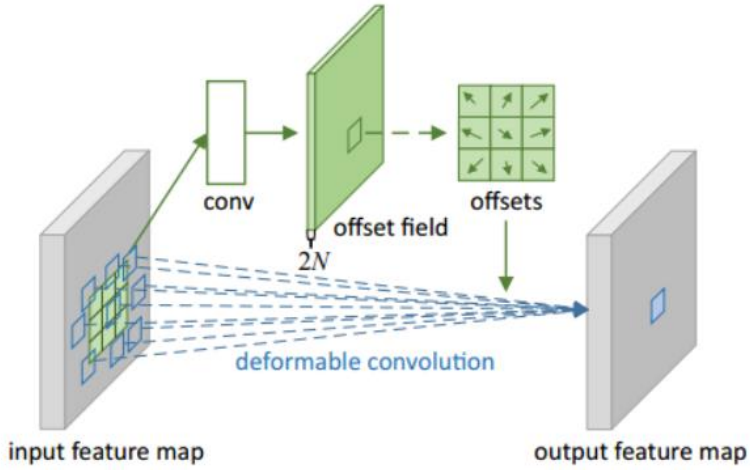
더 넓은 범위에서 grid내의 값을 sampling

object size 와 receptive field의 상관관계 존재

background, large object의 경우 넓은 범위의 receptive field 필요

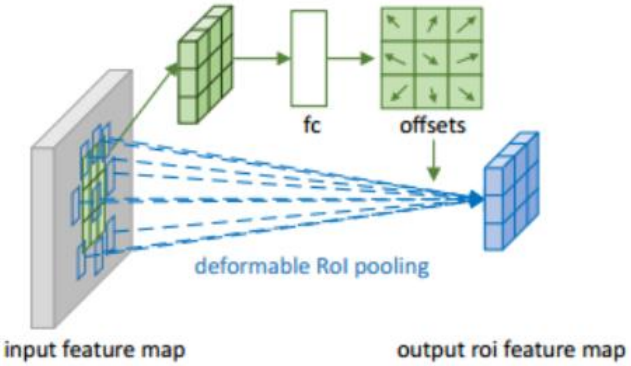
2가지의 방법 제시

deformable convolution



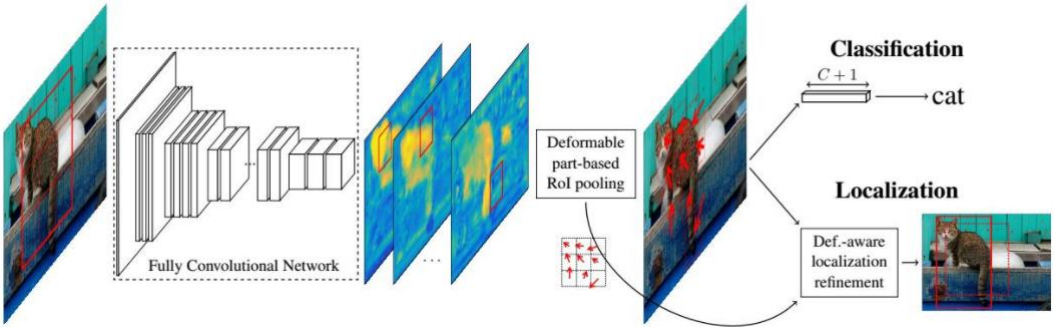
큰 객체 일수록 receptive field가 넓다

deformable RoI pooling



좀더 중요한 정보를 갖고 있는 RoI를 이용함을 확인

Deformable Part-based Fully Convolutional Network (DPFCN)



deformable part-based RoI pooling layer제안

• Occlusions : information loss from object instance

DCN, Adversarial Network

hard to solve

최근 SOTA는 CutMix 이다.

• Image degradations : image noise

insufficient lighting

low quality cameras

image compression

intentional low-cost sensors on edge devices and wearable devices

degrade performance of visual recognition

최근 발전에도 불구하고, 작은 물체에 대한 탐지 정확도는 큰 물체에 비해 훨씬 낮다. 따라서 작은 개체의 탐지는 여전히 개체 탐지의 주요 과제 중 하나로 남아 있습니다. 자율 주행과 같은 특정 애플리케이션은 더 큰 영역 내에서 작은 물체의 존재 여부만 식별하면 되고 정확한 위치 파악은 필요하지 않다.

(4) Context Reasoning

Contextual information은 object detection and recognition에 도움을 줌

(특히 small object, occluded objects, and with poor image quality 일 때)

널리 3가지의 categories로 분류된다.

1. Semantic context
2. Spatial context
3. Scale context

최근 contextual information(cue)를 DCNN-based object detector에 명시하는 연구들이 있으며 이는 2가지의 categories로 조직됨

Global context

Image or scene level contexts 가 객체 탐지의 단서의 역할을 함

(bedroom 은 bed의 존재를 예측)

DeepIDNet

concatenated classification scores(used as contextual features) and

object detection scores(improve detection results)

ION

spatial Recurrent Neural Networks (RNNs)

(전체 이미지에서 contextual information 탐색)

SegDeepM

Markov random field model

(각 detection에서 appearance 와 context의 점수를 측정)

추가적인 feature 가 확대된 object proposals

Local context

근처 물체 사이의 관계를 고려 (물체간 상호작용과 주변 공간)

BBox(different classes), location, scales 등 추론이 필요한 object modeling 은 challenging 함

객체 관계를 명시적으로 modeling 한 연구들

Spatial Memory Network (SMN)

Object relations reasoning을 위해 또다른 CNN에 입력되기 쉬움

메모리를 추가로 업데이트하는 detection을 얻기 위해 이미지와 메모리가 병렬로 처리됨

ORN

Inspired by attention module(NLP)

appearance feature 와 기하학 사이의 상호작용을 통해 objects set을 연산

First Fully end-to-end object

SIN

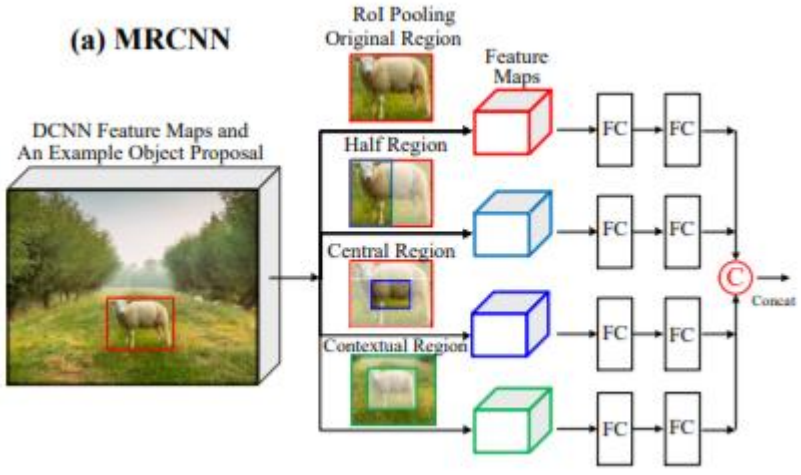
한 이미지에서 두가지의 context를 고려

(scene contextual information, object relationships)

객체 탐지를 그래프 구조 추론으로 공식화합니다. 여기서 객체는 그래프 노드이고 모서리와 관계가 있습니다.

탐지 창 크기를 확대하여 특정 형식의 로컬 컨텍스트를 추출하는 detectors

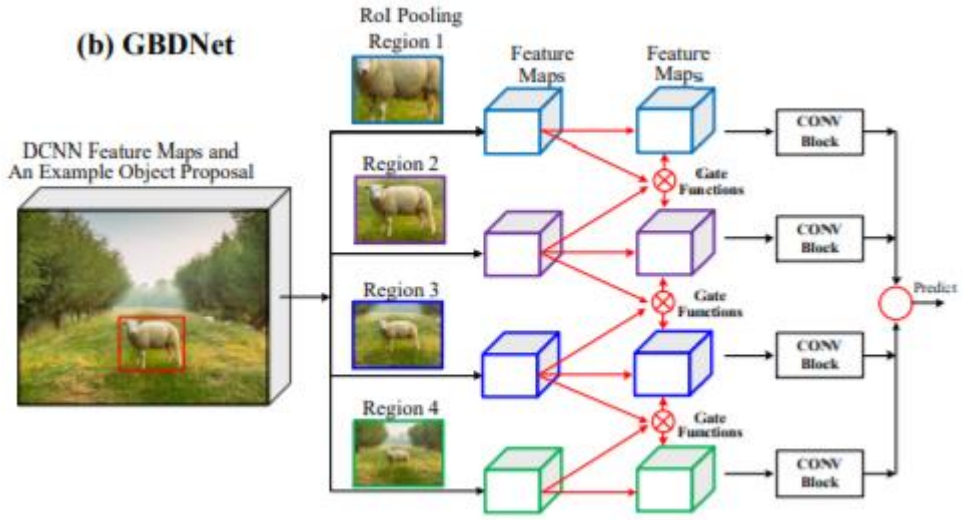
MRCNN



Four contextual regions, 5개의 구조로 되어 있다.

분류기는 전체적으로 여러 경로 따라 훈련

Gated BiDirectional CNN (GBDNet)

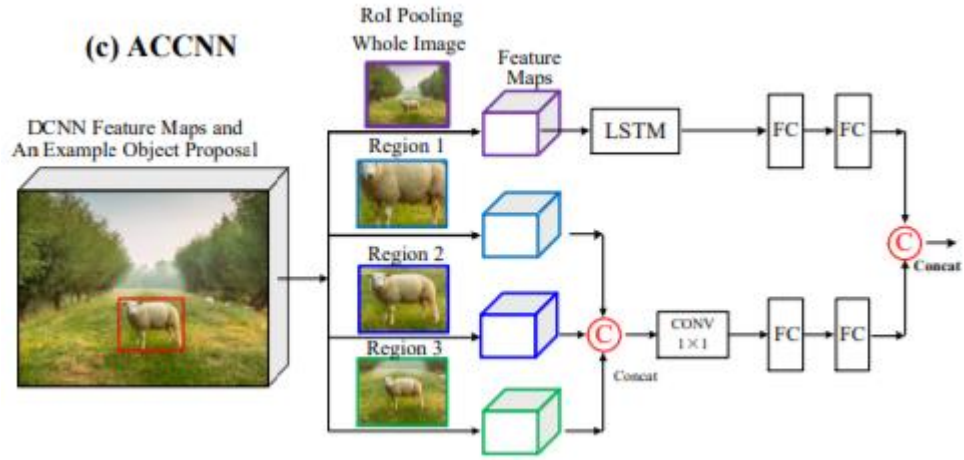


multiscale contextualized regions (object proposal을 감싸는)에서 features 추출

GBDNet은 인접 region proposal 간의 양방향으로 변환을 통해

서로 다른 컨텍스트 영역의 기능 간에 메시지를 전달합니다.

Attention to Context CNN (ACCNN)

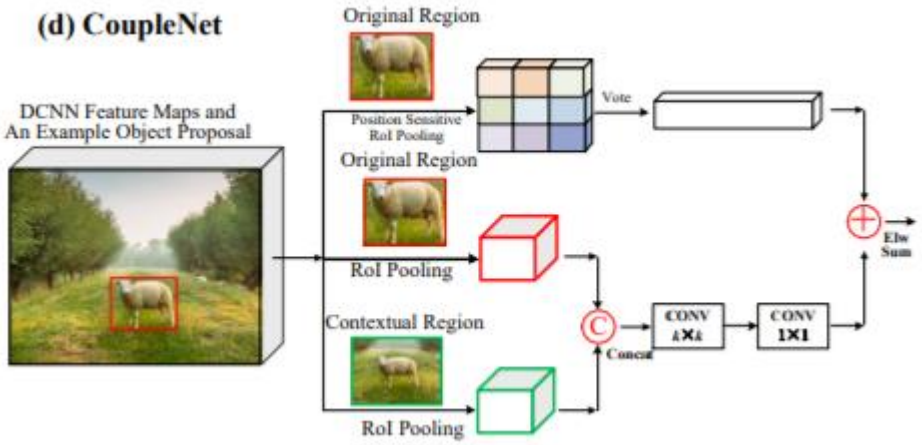


Multiscale Local Contextualized subnetwork (capture global context)

MRCNN (capture local context)

local and global context features are concatenated for recognition

CoupleNet



similar to ACCNN

RFCN 새로운 branch로 생성 (object information을 capture)

branch내 의 RoI pooling (encode the global context)

(5) Detection Proposal Methods

Handcrafted feature descriptors : SIFT, HOG

기존 object detection DPM (Sliding window)의 한계 :

image의 pixel이 증가함에 따라 window size크기 증가하고

search하는 영역이 넓어짐에 따라 계산량 증가

Region proposals (object를 담는 region의 집합) 제안

적절 개수의 region proposals(Pre-processing step)을 통해 high object recall 달성 가능

(기존의 sliding windows approach 보다 빠른 성능, 정교한 classifiers 사용 가능)

Pre-processing step 의 특징

1. High recall
2. Accurate localization
3. Low computational cost

2014년 DNN features와 region proposals를 결합하여 object detection 성능 향상

Approach base 변화

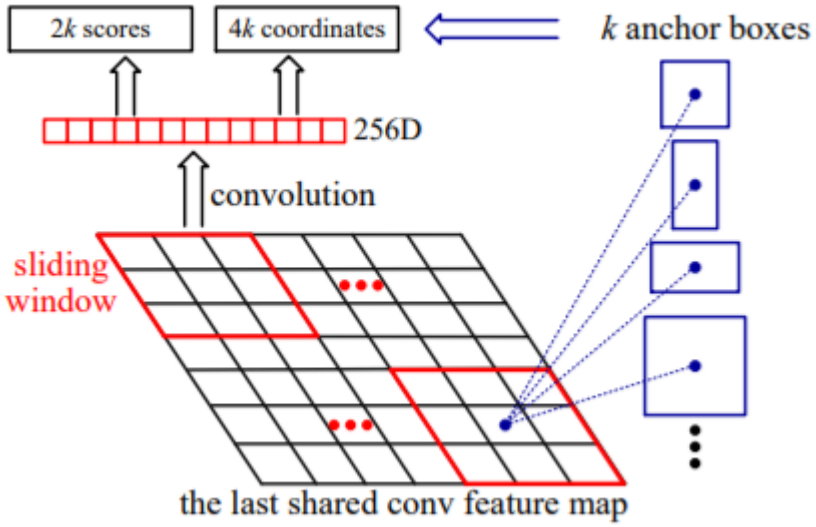
1. Traditional low-level cues (color, texture, edge, gradients)
2. Selective Search, MCG, EdgeBoxes
3. DCNN

DCNN based object proposal method 는 2가지의 categories로 분류

Bounding Box Proposal Methods

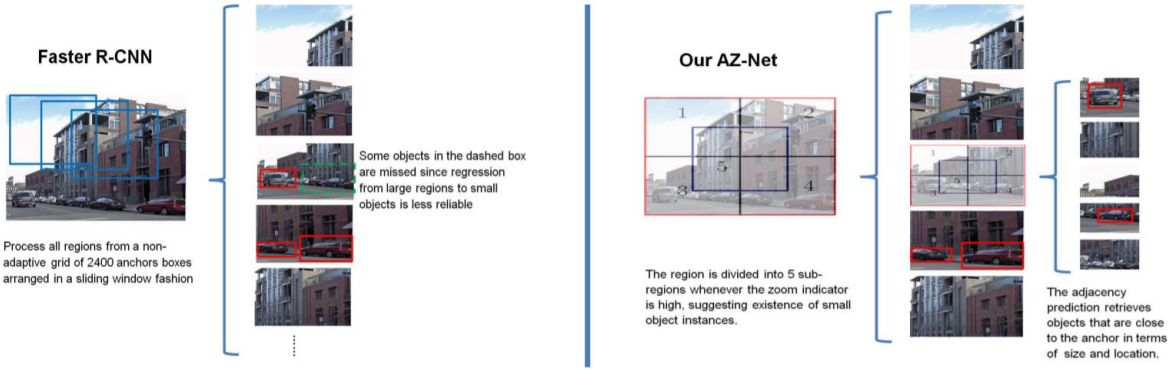
RPN

전체 이미지를 convolution 한 후의 feature을 detection에 공유 (처음)



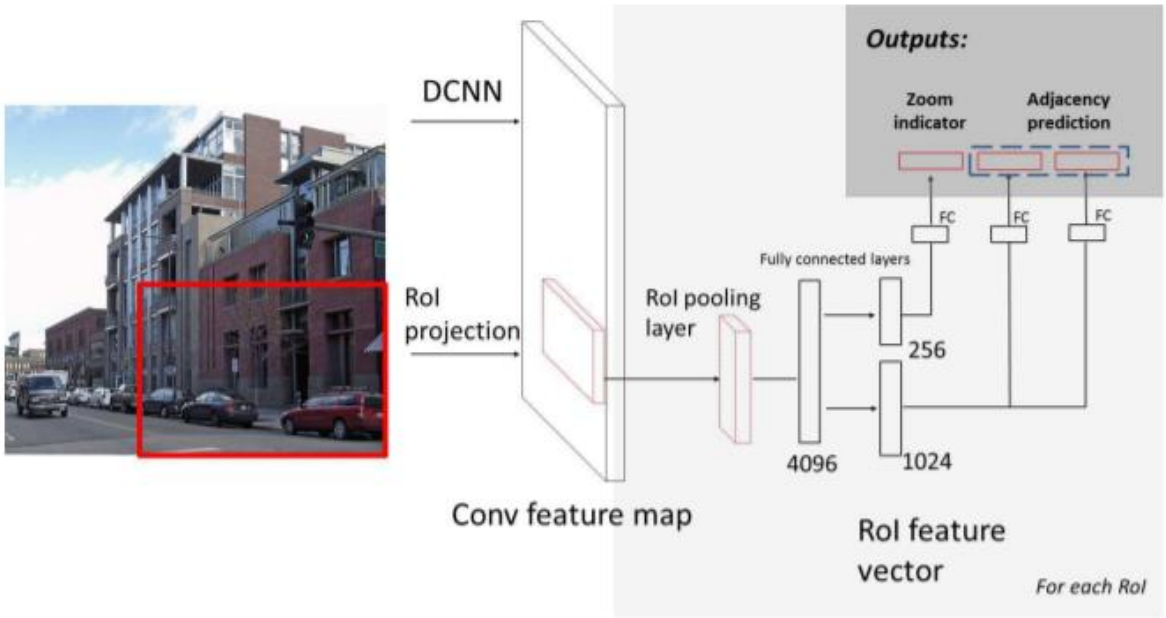
sliding window를 사용, k개의 anchor box를 통해 k개의 proposals 예측

AZNet



Coarse-to-fine search

(큰 region에서 시작하여 반복적으로 object가 포함된 subregion을 찾는다)



RPN에서 scalar zoom indicator를 계산하는 branch 추가

Exploiting multilayer convolutional features

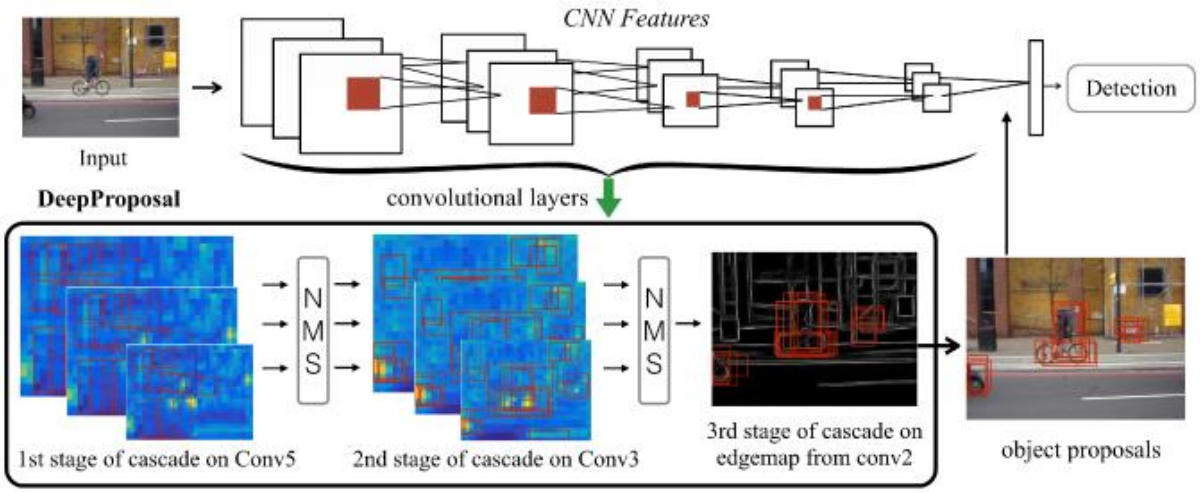
DeepBOX

lightweight CNN 사용

EdgeBox로 생성된 proposals를 rerank하기 위해 학습

Detection에 feature를 공유하지 않음

DeepProposal



multilayer convolutional features로 inverse cascade를 build

coarse-to-fine 방법으로 region proposal 획득

HyperNet

Hyper Features (multilayer convolution features를 집계)

Shares in generating proposals

end-to-end joint training strategy

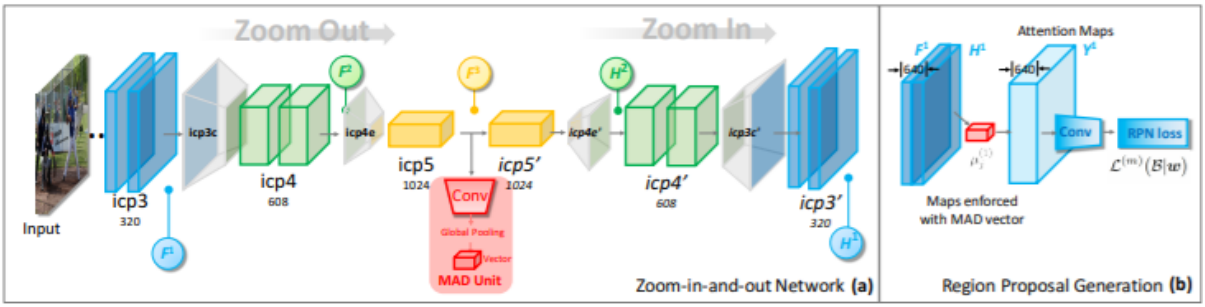
CRAFT

cascade strategy

(RPN을 학습하여)region proposals를 생성

생성된 것을 binary Fast RCNN network에 학습 (object 와 background를 분리)

Zoom Out-and-In Network (ZIP)



MAD(Map Attention Decision Unit)

저수준 및 고수준 스트림의 피쳐 맵 중 뉴런 활성화를 적극적으로 검색

low level detail과 high level semantic를 통합

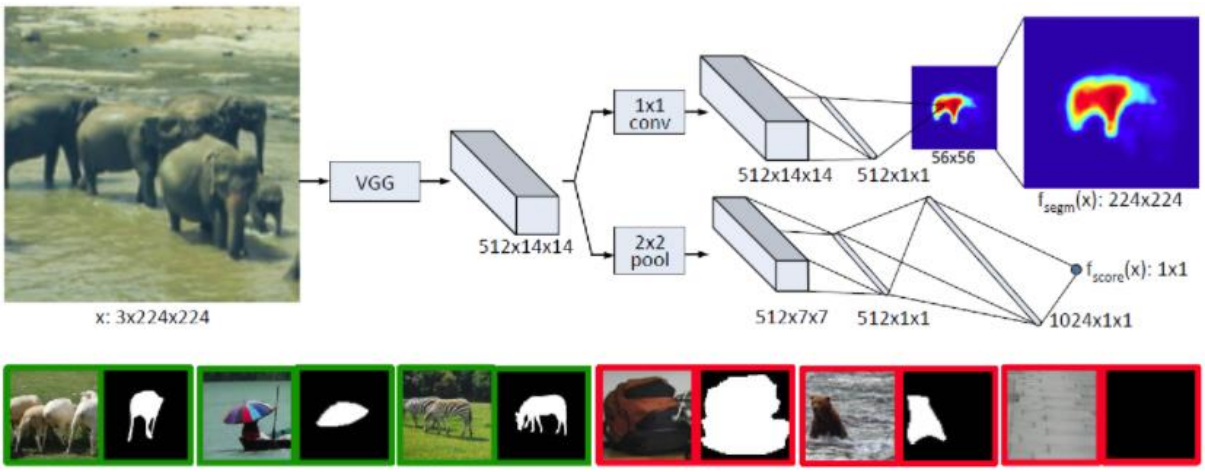
DeNet

BBox corner estimation 을 하여 region proposals 예측 (RPN 대체)

Object Segment Proposal Methods

BBox proposals 보다 더욱 informative하다.

DeepMask



처음 DCNN으로 object mask proposals 을 생성한 연구

학습 시 Segment Annotation 필요

OverFeat

feature map에 convolution filter를 적용하여 전체순회

fc layer를 convolution layer로 대체 (sliding window와 같은 효과)

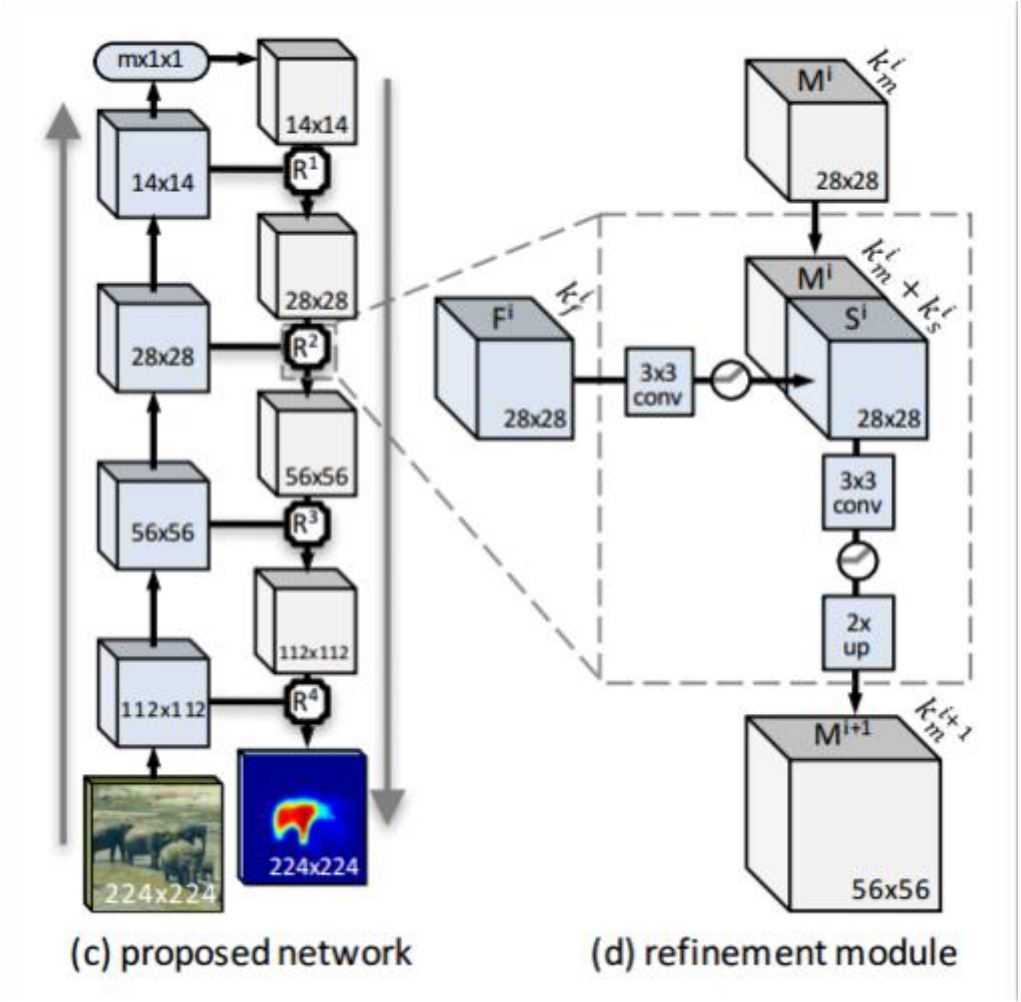
InstanceFCN

FCN과 DeepMask를 결합

2개의 fully convolutional branches

(instance sensitive score map, 물체 판별 score)

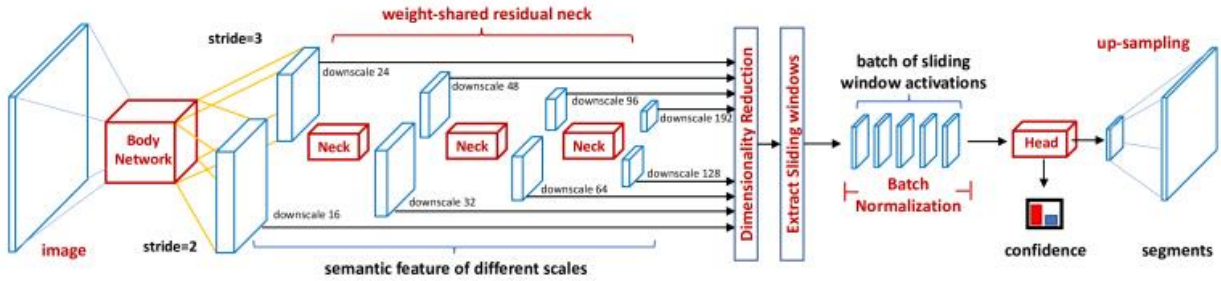
SharpMask



Generate high fidelity(precise) mask

(fusion of low-level features and high level features)

Fast-Mask



SSD와 유사한 One-Shot manner

(6) Other Factors

Data augmentation

Novel training strategies

combinations of backbone models

Multiple detection frame works

Incorporating information from other related tasks

Methods for reducing localization error

Handling the huge imbalance between positive and negative samples

Improving loss functions