OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks

arXiv 2013, 4314 citation

GIST 20205035 김연혁

Three purpose of the paper

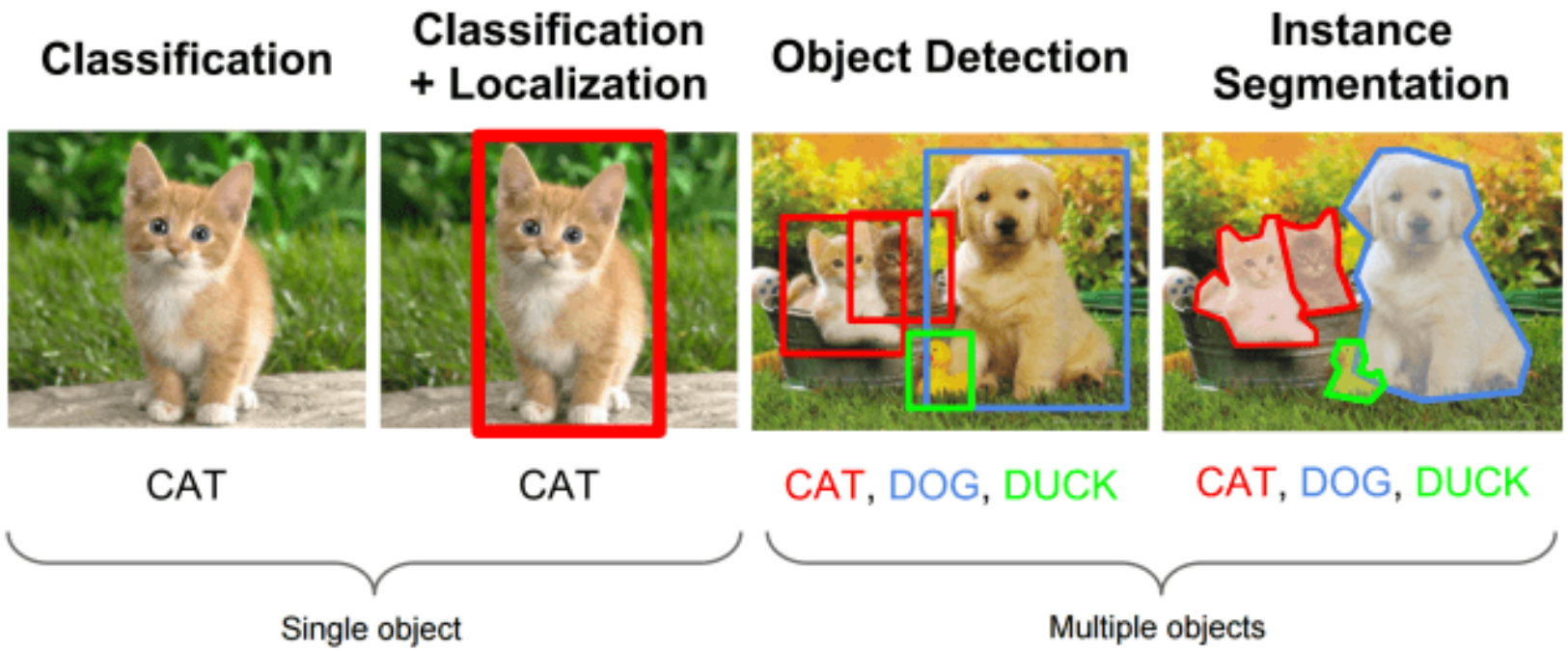
Recognition을 사용한 Localization과 Detection Task의 새로운 딥러닝 접근법 제안

CNN 기반 Recognition, Localization, Detection 통합 framework 제안

개별 Task를 학습하는 것보다 세 Task를 동시에 학습하는 것의 성능이 향상됨

Recognition, Localization, Detection

Detection = Recognition + Localization



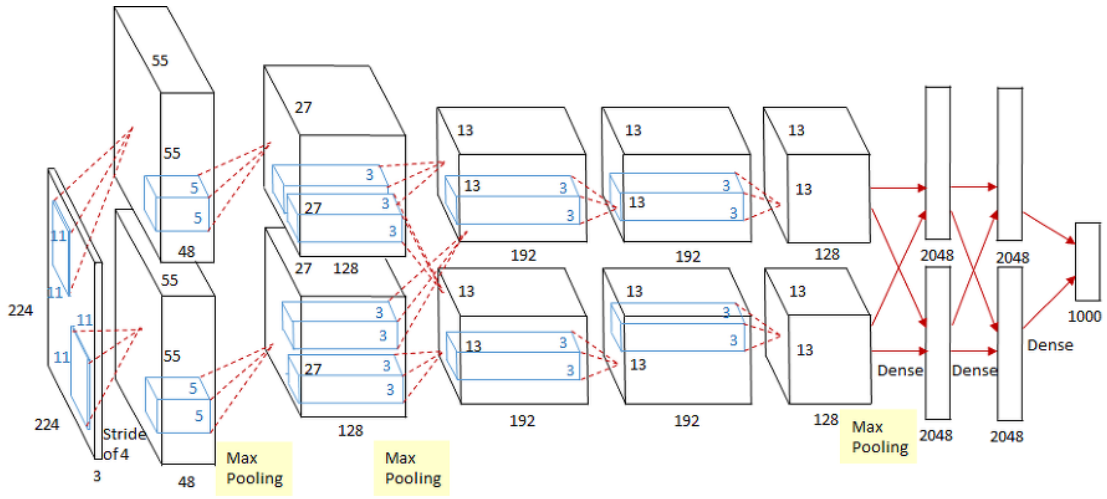
Stanford University School of Engineering

Main Architecture of Overfeat

Almost same with AlexNet by Krizhevsky from "Imagenet classification with DCNN"

Image -> 256 x 256 -> 221 x 221
downsampling random crop

Adjust of num of stride,
batch size,
pooling region (겹치지 않도록)



Layer	1	2	3	4	5	6	7	Output 8
Stage	conv + max	conv + max	conv	conv	conv + max	full	full	full
# channels	96	256	512	1024	1024	3072	4096	1000
Filter size	11x11	5x5	3x3	3x3	3x3	-	-	-
Conv. stride	4x4	1x1	1x1	1x1	1x1	-	-	-
Pooling size	2x2	2x2	-	-	2x2	-	-	-
Pooling stride	2x2	2x2	-	-	2x2	-	-	-
Zero-Padding size	-	-	1x1x1x1	1x1x1x1	1x1x1x1	-	-	-
Spatial input size	231x231	24x24	12x12	12x12	12x12	6x6	1x1	1x1

Training Detail

Dataset: ImageNet 2012 training set, 1.2 million images, 1000 classes

Image -> downsampling -> 256 x 256 pixels -> 5 random crops –> 221 x 221 pixels

128 mini batch size

Weight initialize $(\mu, \sigma) = (0, 1 * 10^{-2})$

Weight update with Stochastic gradient descent

Momentum term 0.6, l_2 weight decay of $1 * 10^{-5}$

Learning rate init = $5 * 10^{-2}$, 0.5배씩 감소

Deference with AlexNet

No contrast normalization is used

Pooling regions are non-overlapping

Has larger 1st and 2nd layer feature maps
  Smaller stride (2 instead of 4)
  Larger stride is beneficial for speed but will hurt accuracy

Multi view voting -> Dense evaluation

Dense Evaluation

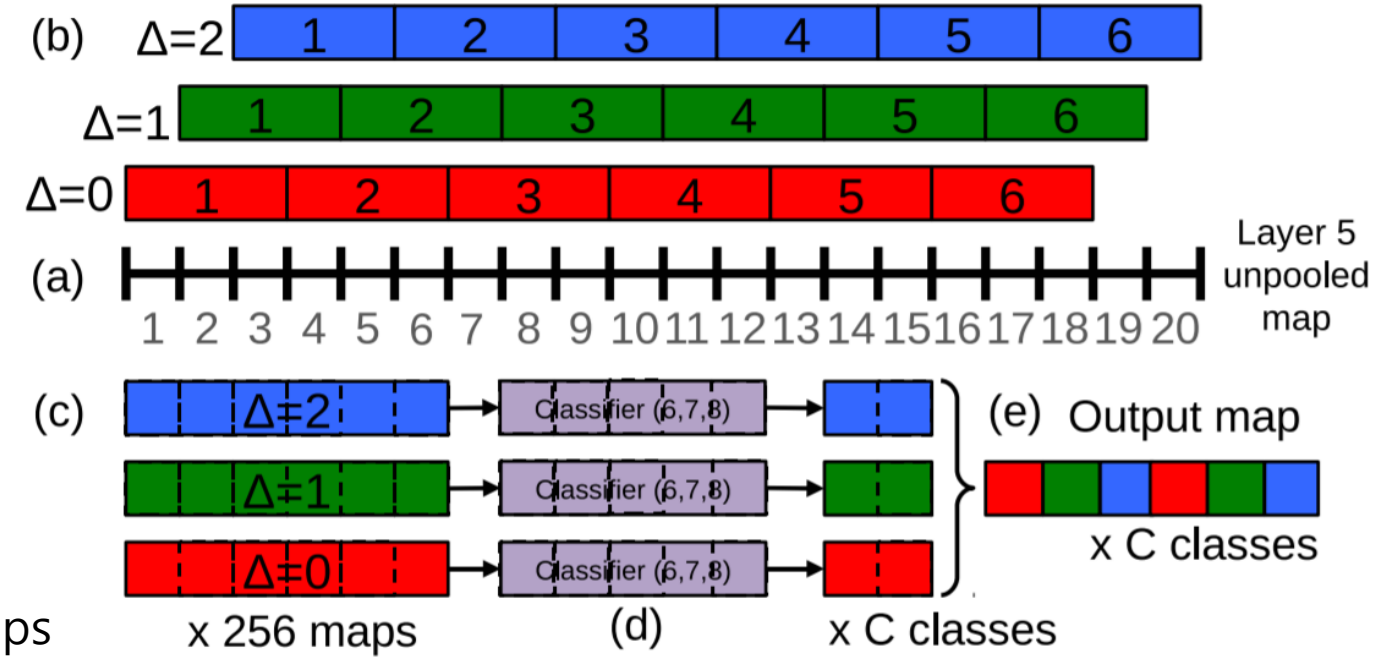
(a): 최종 max-pooling layer

(b): 3x1 non-overlapped
pooling with offset

(c): Resulting 6pixel pooled maps

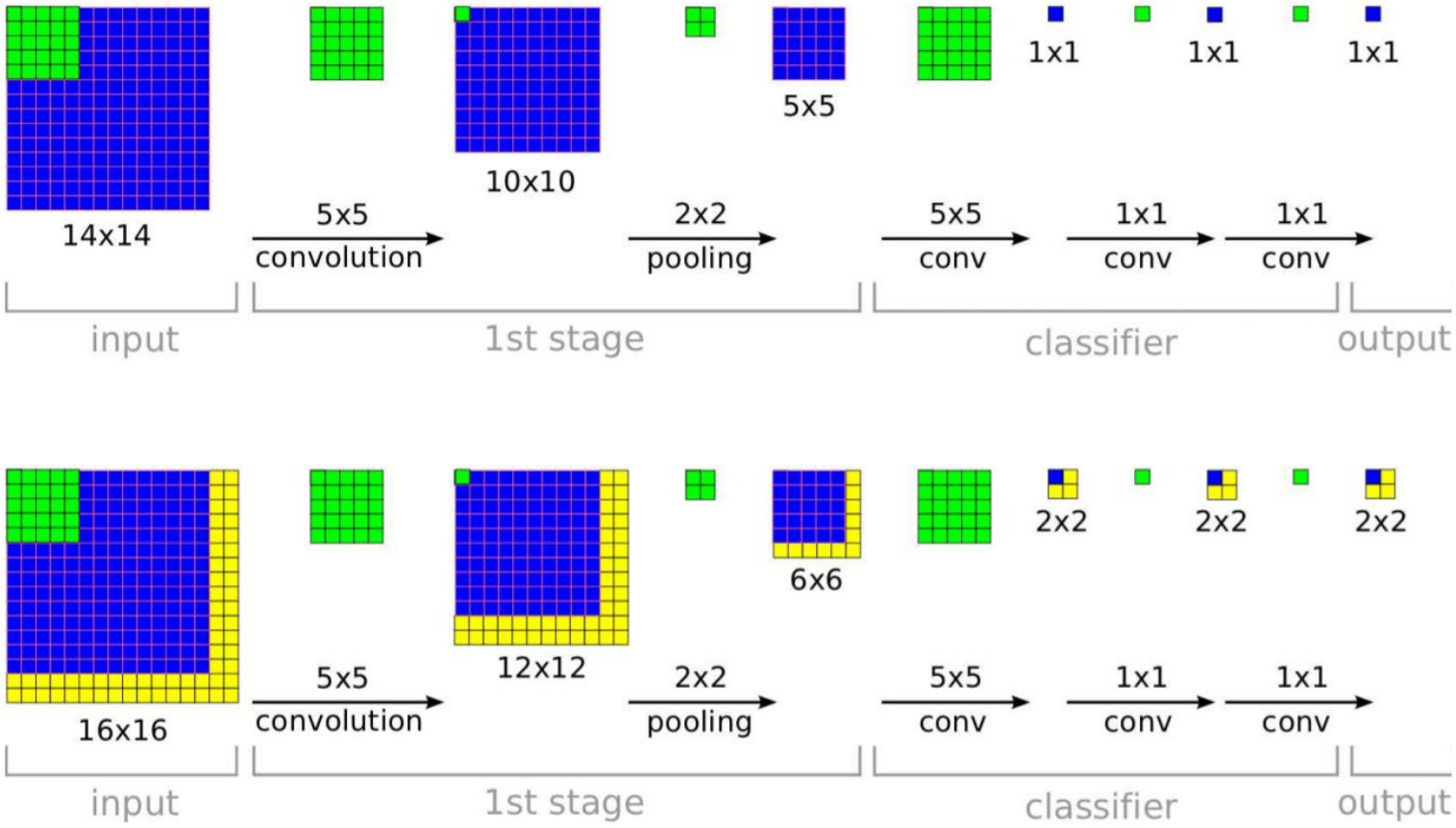
(d): Classifier

(e): Reshaped into 6 pixel by C output maps



Pooling을 통해 resolution이 1/3으로 줄어드는 것을 offset을 1을 주면서
pooling을 진행하면 pooling 이전의 해상도를 유지할 수 있음, 보다 조밀한 검사를 할 수 있음

Dense Evaluation



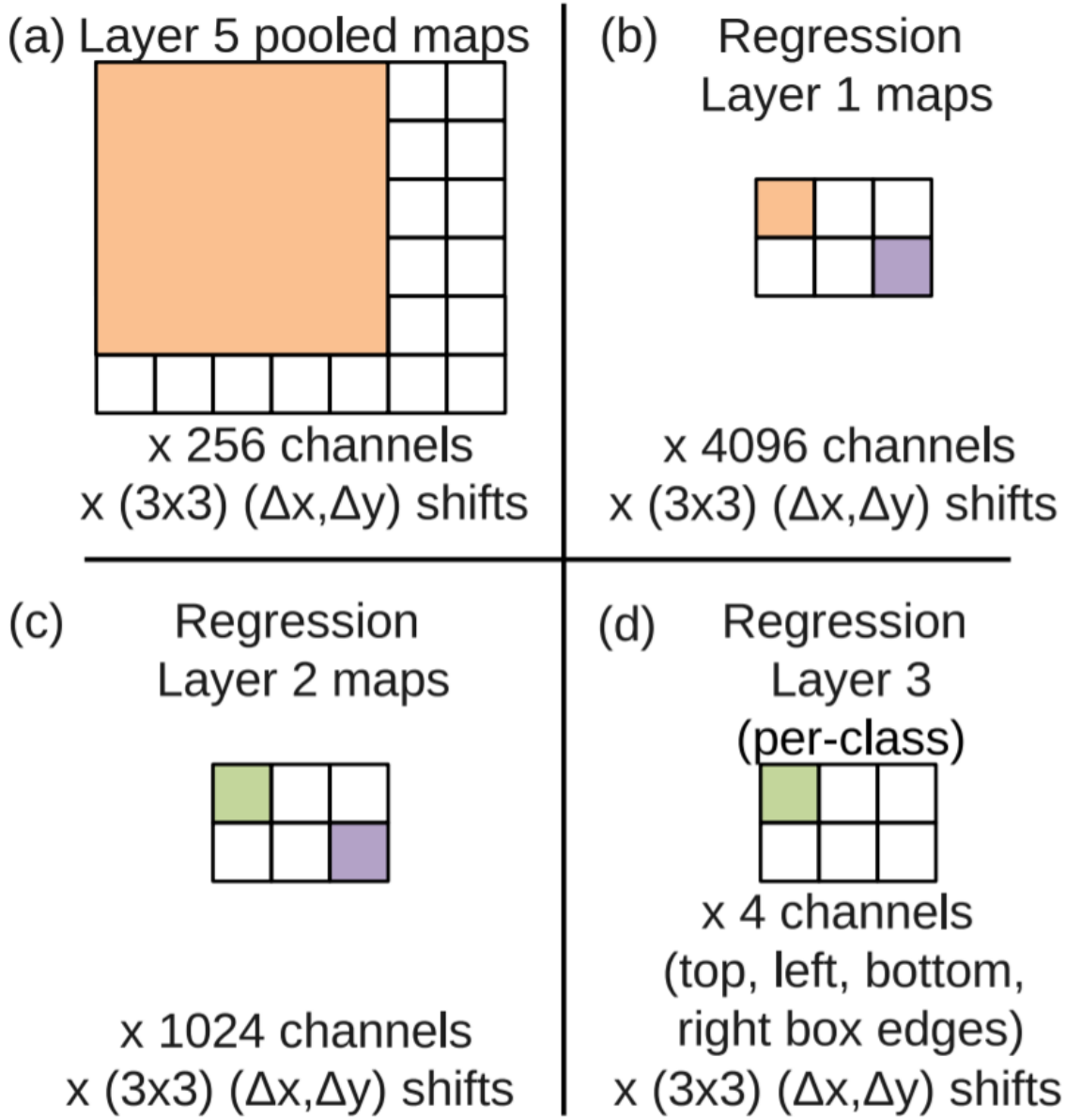
Localization

Recognition을 학습한 모델을 가져옴

Classifier layer -> regression network

(a): 최종 max-pooling layer

(d)에 나온 bounding box 정보를 target으로 하여 학습



Localization

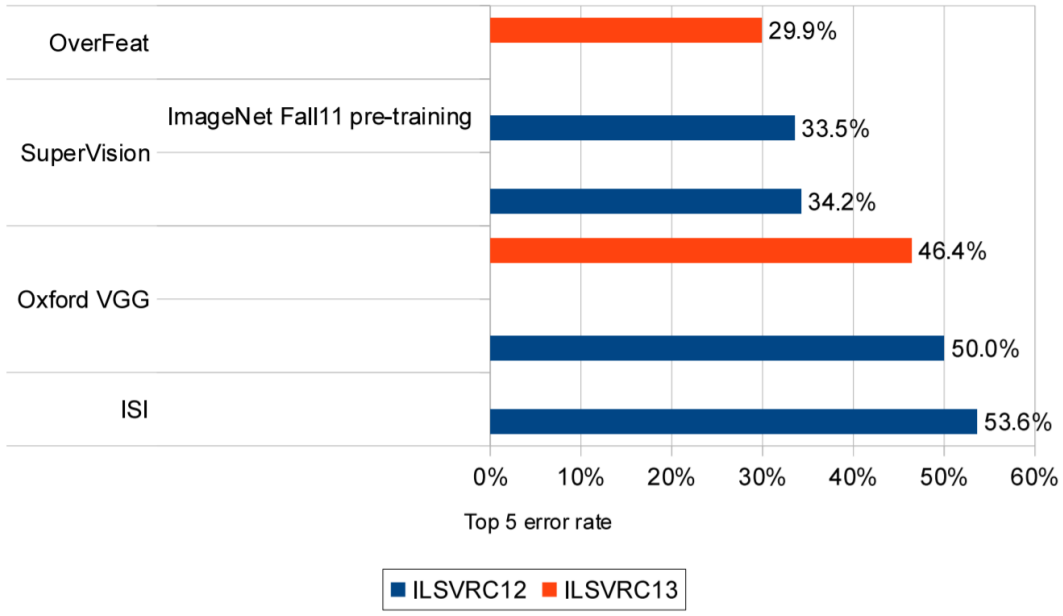
match_score: sum of the distance between centers of the two bounding boxes and the intersection area of the boxes

confidence_score: output of the final softmax layer for a class c

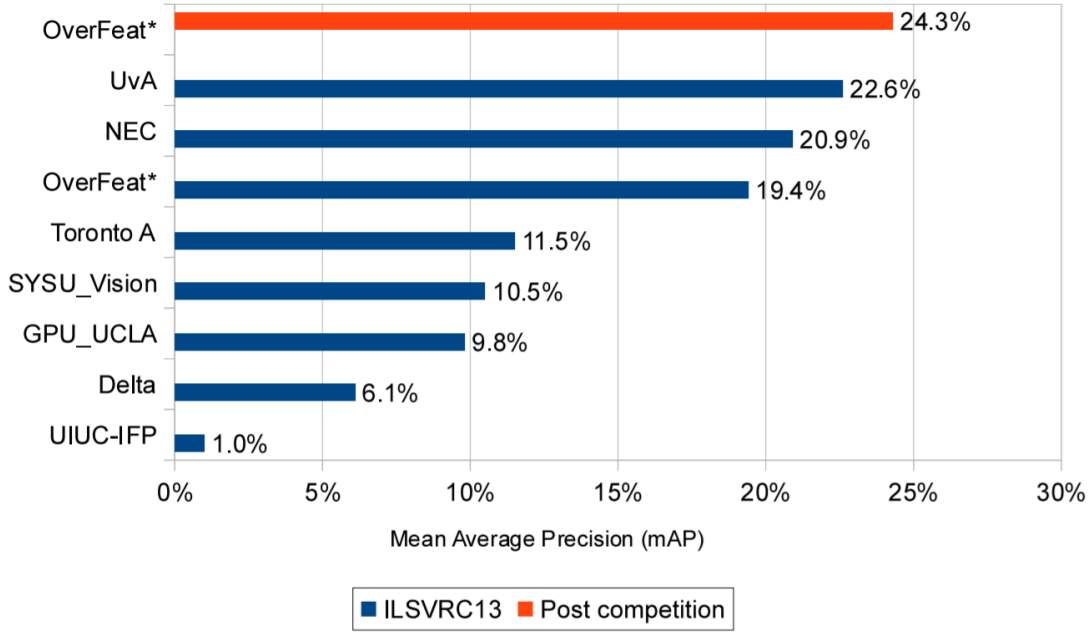
Detection & Result

Localization에서 학습된 weight를 공유

아무 object도 없는 이미지도 같이 학습



Localization



Detection

Conclusions

Recognition, Localization, Detection 세 분야 모두 우수한 성적을 보인 모델은 유일

Recognition, Localization, Detection 세 Tasks를 weight를 공유해가며 학습하며 성능을 향상시킴

Dense evaluation을 사용하여 불필요하게 중복되는 연산을 줄임

추가로 참고한 자료의 출처
라온피플㈜ 블로그 https://blog.naver.com/laonple/220752877630
wujincheon github blog https://wujincheon.github.io/wujincheon.github.io/deep%20learning/2019/02/15/overfeat.html