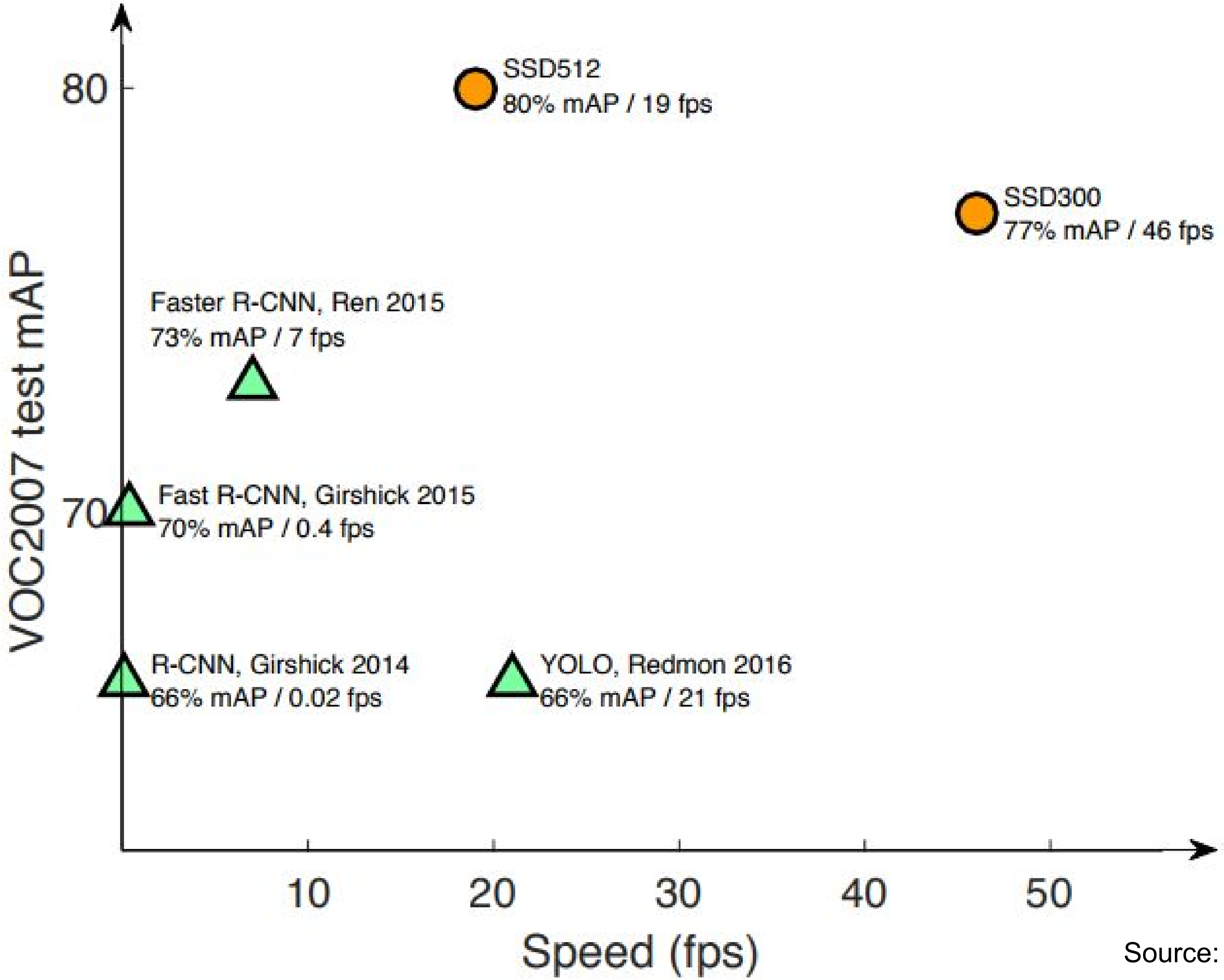
SSD:Single Shot MultiBox Detector

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Redd, Cheng-Yang Fu, Alexander C.Berg

Presenter: Hongjing Zhang, Chen Zhang

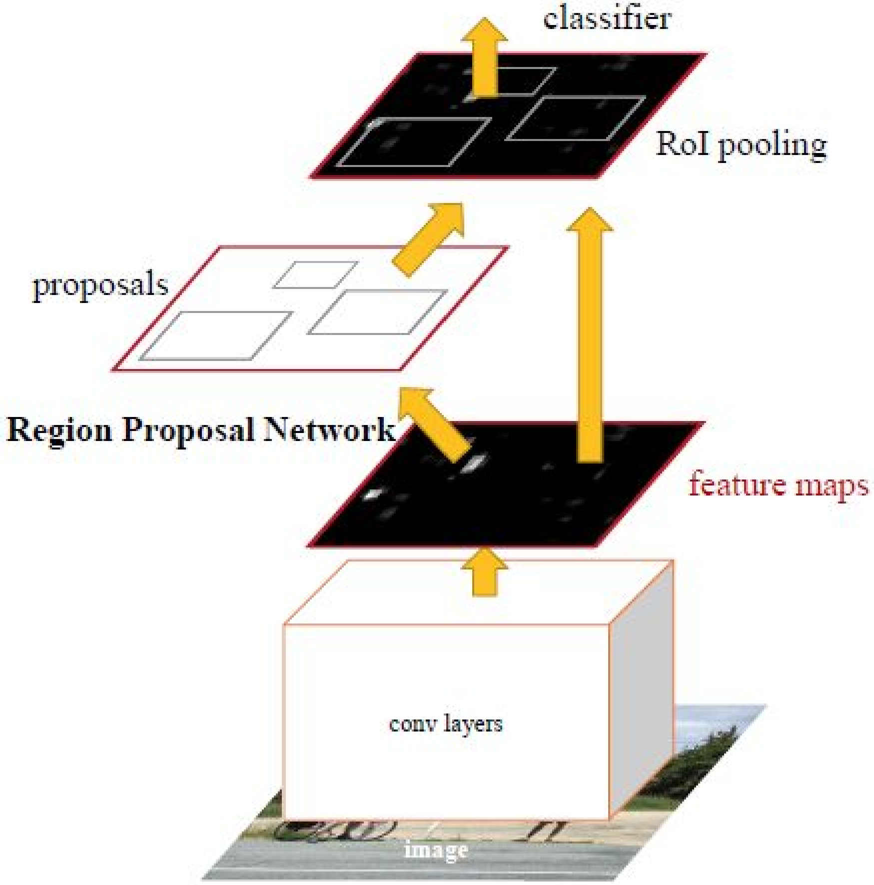
Table of contents

Performance On VOC2007

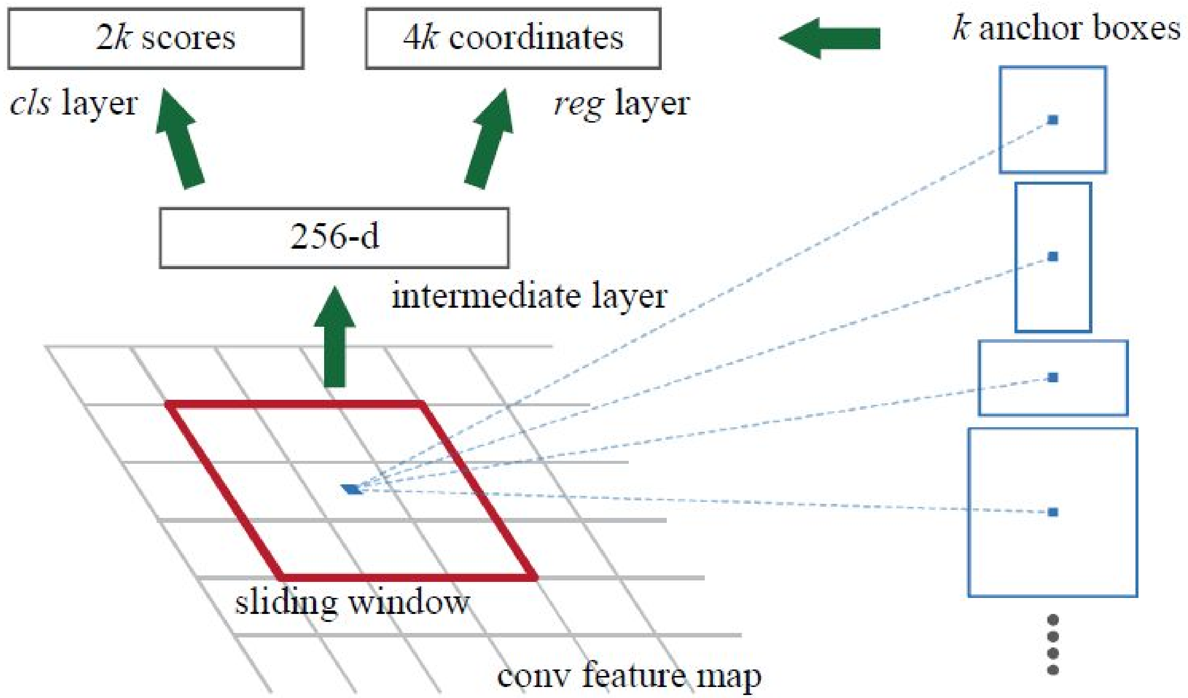


Source: http://www.cs.unc.edu/~wliu/papers/ssd_eccv2016_slide.pdf

Quick Review of Faster R-CNN



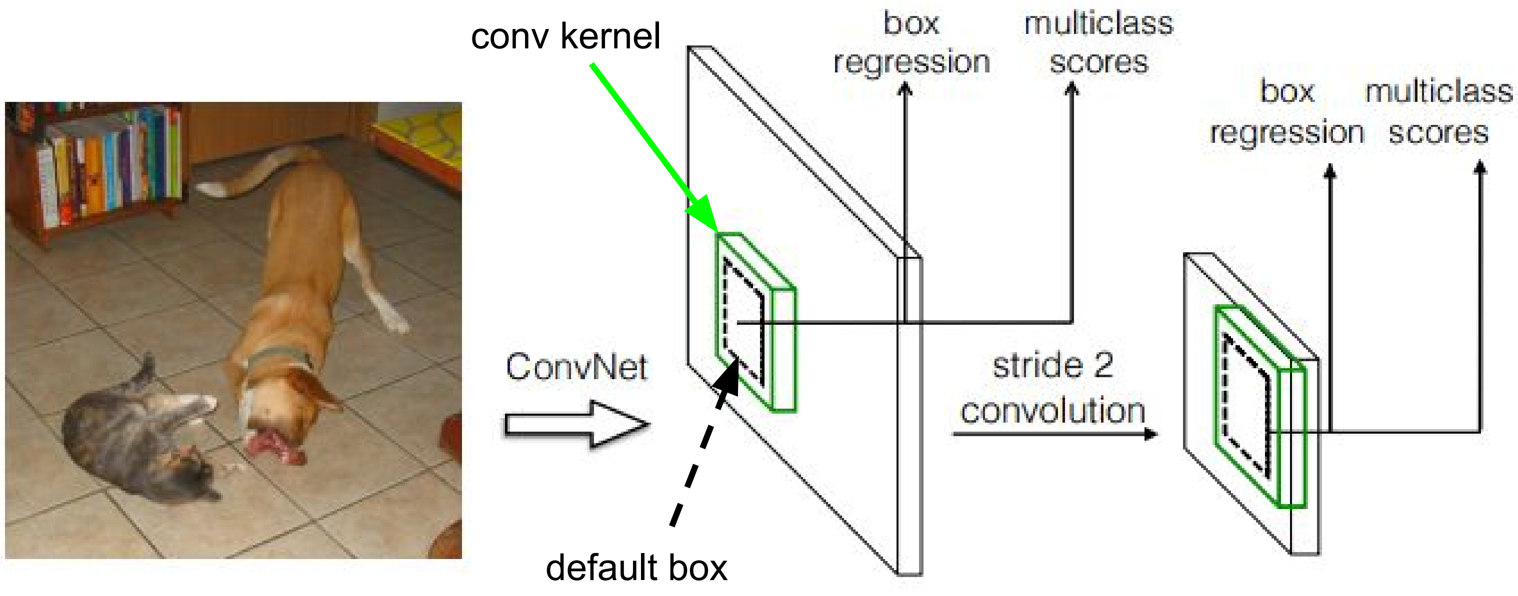
Fast R-CNN + Region Proposal Net



Region Proposal Net: k anchor boxes of different scale and aspect ratio at each position.

source: faster R-CNN paper by S. Ren

SSD: Multi-scale Feature Map



Feature maps from different conv layers of different sizes.

Source: http://www.cs.unc.edu/~wliu/papers/ssd_eccv2016_slide.pdf

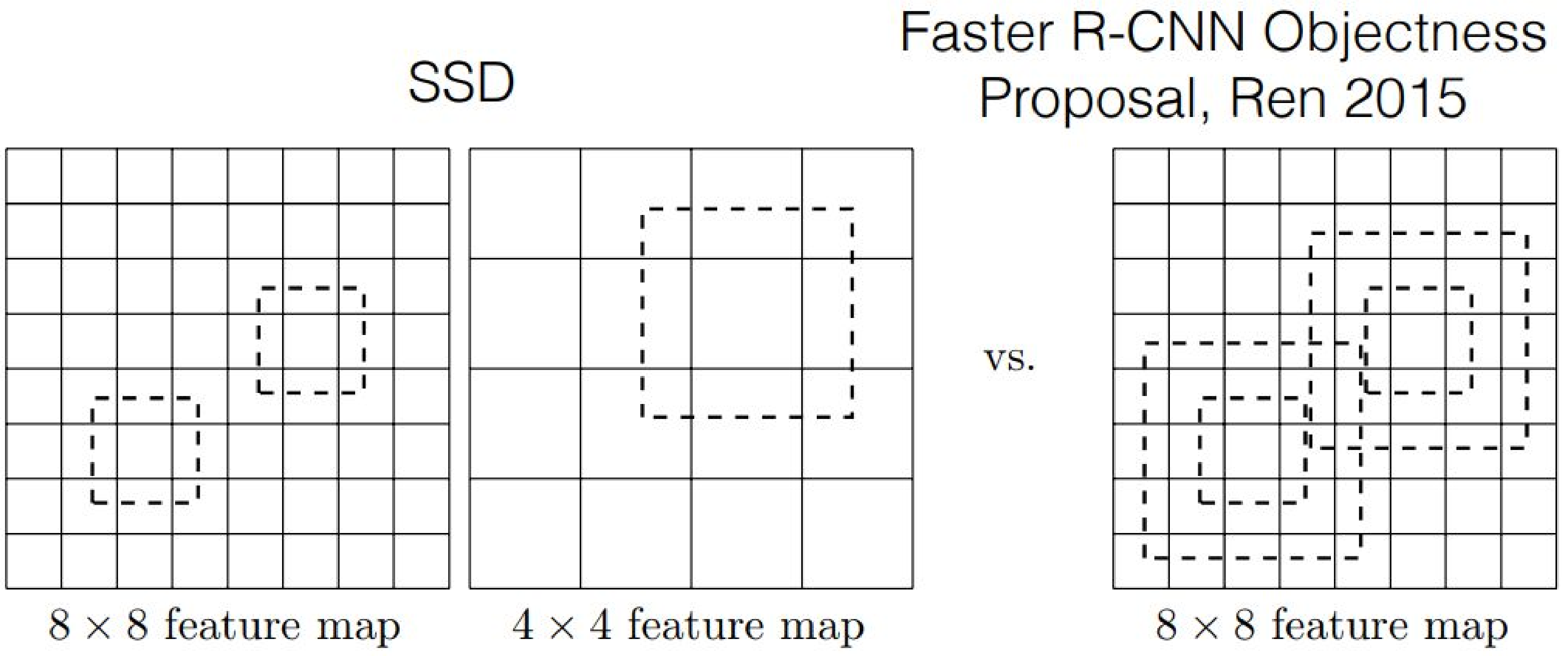
SSD: Multi-scale Feature Map

Prediction source layers from:						mAP use boundary boxes?	
38×38	19×19	10×10	5×5	3×3	1×1	Yes	No
✔	✔	✔	✔	✔	✔	74.3	63.4
✔	✔	✔				70.7	69.2
	✔					62.4	64.0

Effects on mAP when using different feature maps.

Source: http://www.cs.unc.edu/~wliu/papers/ssd_eccv2016_slide.pdf

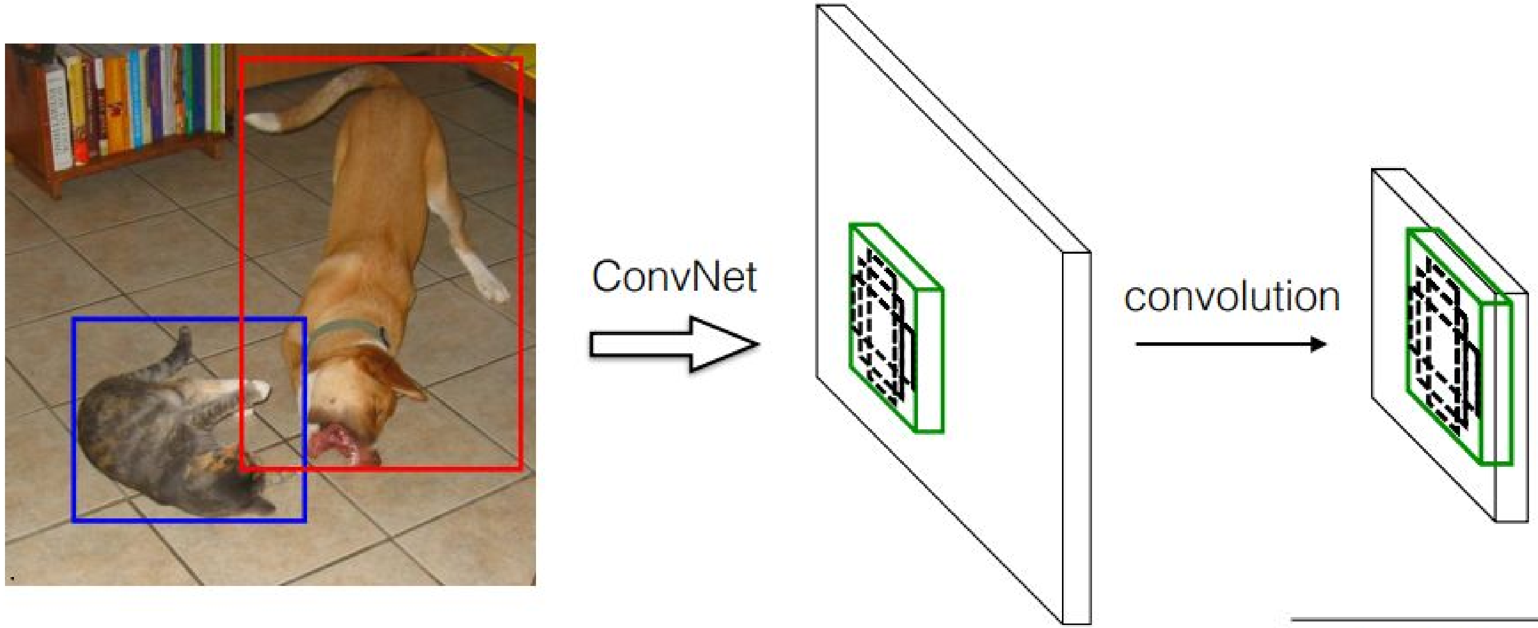
SSD: Multi-scale Feature Map



Anchor box setting comparison: SSD vs. Faster R-CNN

Source: http://www.cs.unc.edu/~wliu/papers/ssd_eccv2016_slide.pdf

SSD: Multi-scale Feature Map



Anchor box setting: aspect ratios

	SSD300		
include $\{\frac{1}{2}, 2\}$ box?		✔	✔
include $\{\frac{1}{3}, 3\}$ box?			✔
number of Boxes	3880	7760	**8732**
VOC2007 test mAP	71.6	73.7	**74.3**

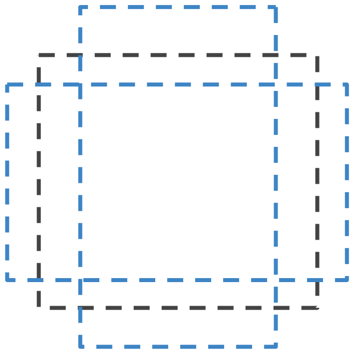
Source: http://www.cs.unc.edu/~wliu/papers/ssd_eccv2016_slide.pdf

Default Bounding Boxes - scale and shape

Aspect ratio: *a* in $\{1, 2, 3, \frac{1}{2}, \frac{1}{3}\}$ One extra default case: $s'_k = \sqrt{s_k s_{k+1}}$

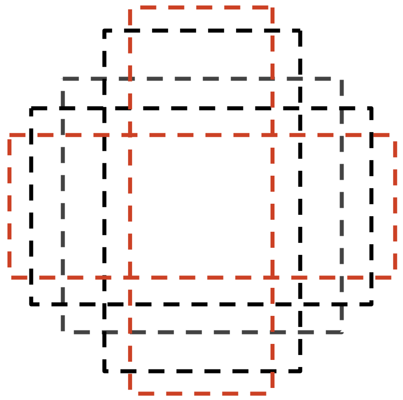
Default Bounding Boxes - scale and shape

Aspect ratio: *a* in $\{1, 2, 3, \frac{1}{2}, \frac{1}{3}\}$ One extra default case: $s'_k = \sqrt{s_k s_{k+1}}$



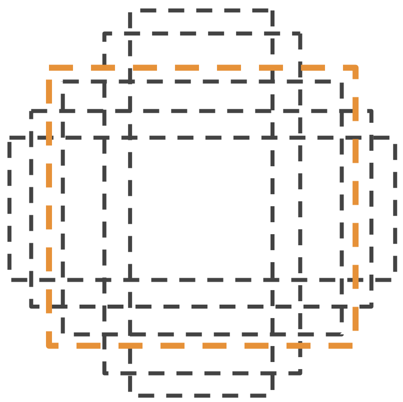
Default Bounding Boxes - scale and shape

Aspect ratio: *a* in $\{1, 2, 3, \frac{1}{2}, \frac{1}{3}\}$ One extra default case: $s'_k = \sqrt{s_k s_{k+1}}$



Default Bounding Boxes - scale and shape

Aspect ratio: *a* in $\{1, 2, 3, \frac{1}{2}, \frac{1}{3}\}$ One extra default case: $s'_k = \sqrt{s_k s_{k+1}}$



Default Bounding Boxes - scale and shape

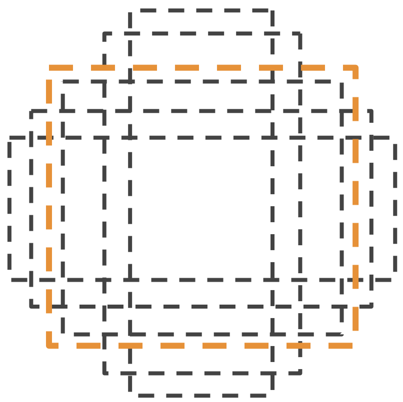
Aspect ratio: *a* in $\{1, 2, 3, \frac{1}{2}, \frac{1}{3}\}$ One extra default case: $s'_k = \sqrt{s_k s_{k+1}}$

Scale of default boxes computed as a linear function:

$$s_k = s_{\text{min}} + \frac{s_{\text{max}} - s_{\text{min}}}{m-1}(k-1), \quad k \in [1, m]$$

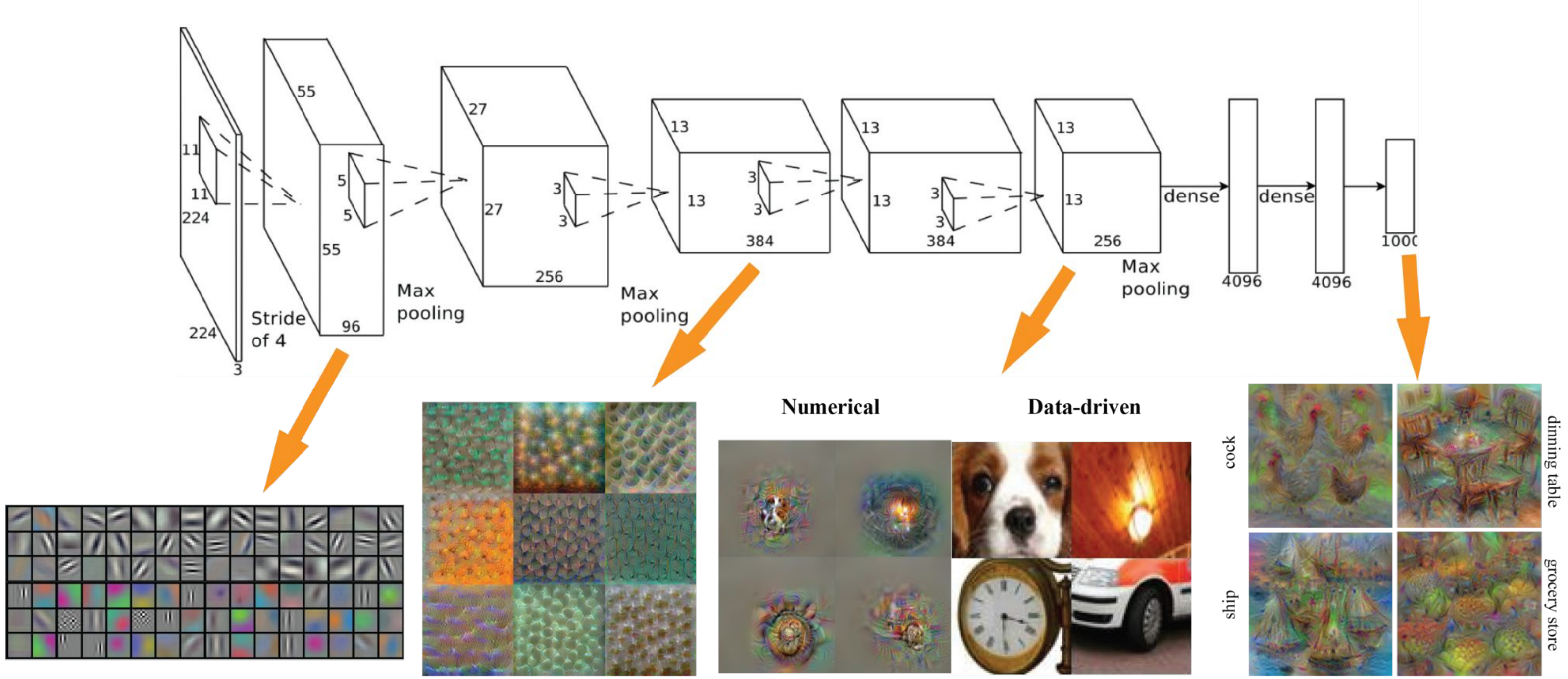
*S*min = 0.2, *S*max = 0.9

width ($w_k^a = s_k\sqrt{a_r}$) and height ($h_k^a = s_k/\sqrt{a_r}$)



Default Bounding Boxes

Why small boxes in large feature maps?



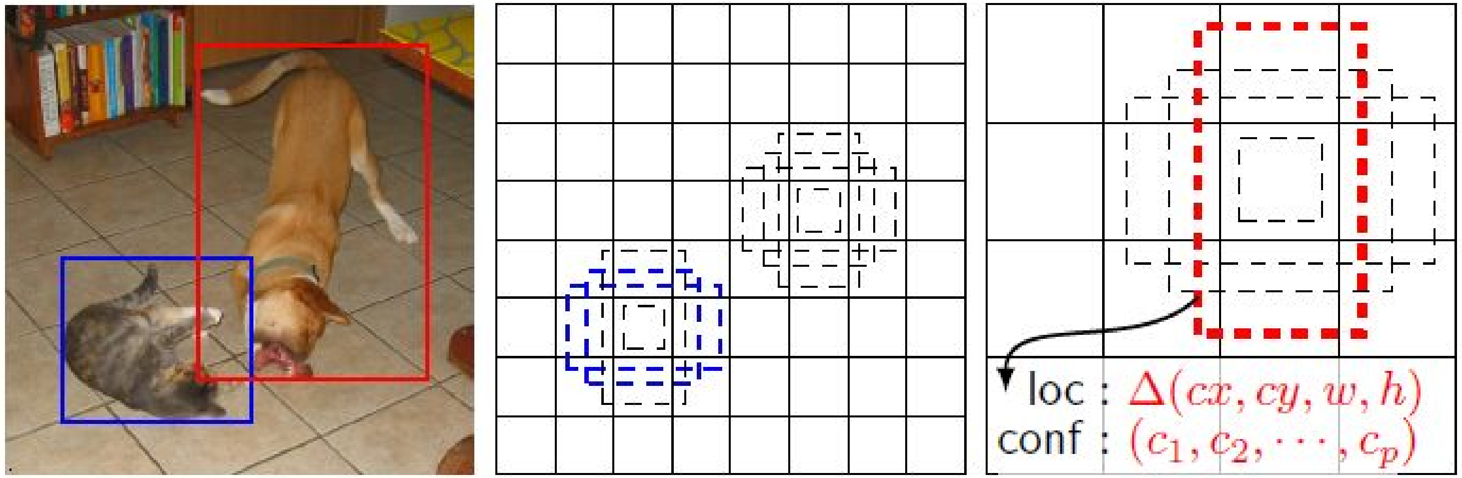
Conv 1: Edge+Blob Conv 3: Texture Conv 5: Object Parts Fc8: Object Classes

Source: https://towardsdatascience.com/understanding-ssd-multibox-real-time-object-detection-in-deep-learning-495ef744fab

Default Bounding Boxes

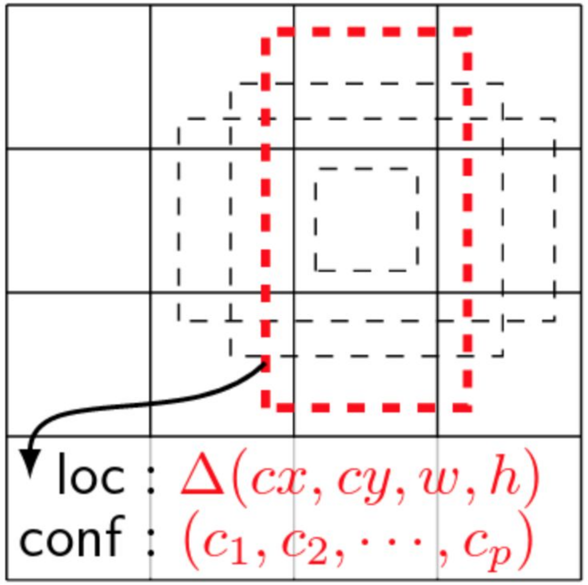
Why small boxes in large feature maps?

- large feature map - small receptive field - small object
- small feature map - large receptive field - large object



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

Convolutional predictors for detection



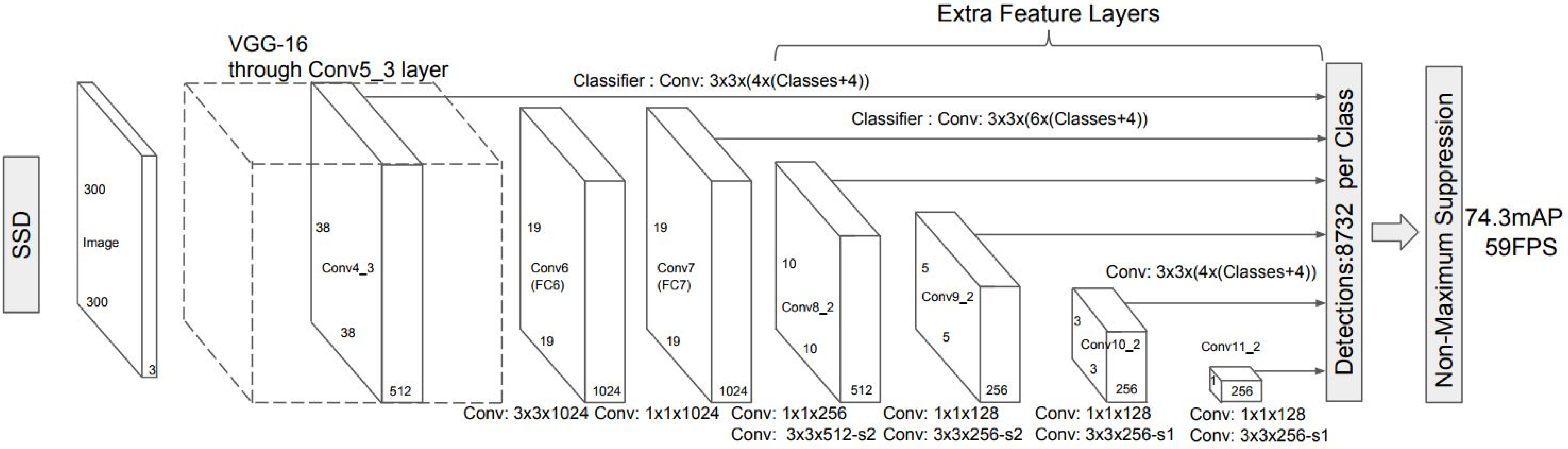
3 x 3 x p kernel

Each box:

- cls: # classes (C1, C2, …, Cp)
- reg: 4 parameters delta(cx, cy, w, h)

conv kernels: (Classes+4) x (# Default Boxes)

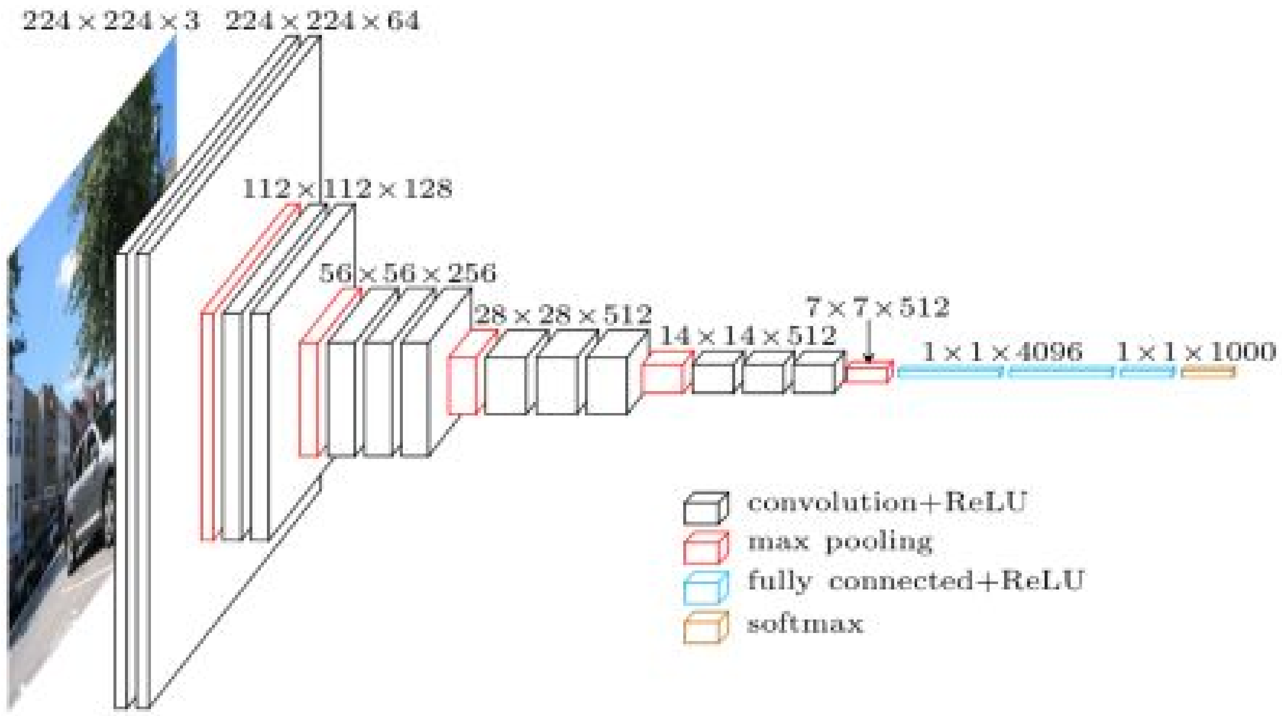
SSD Network Structure



SSD architecture taken from the original paper

VGG-16 network: by Oxford's Visual Geometry Group (VGG)

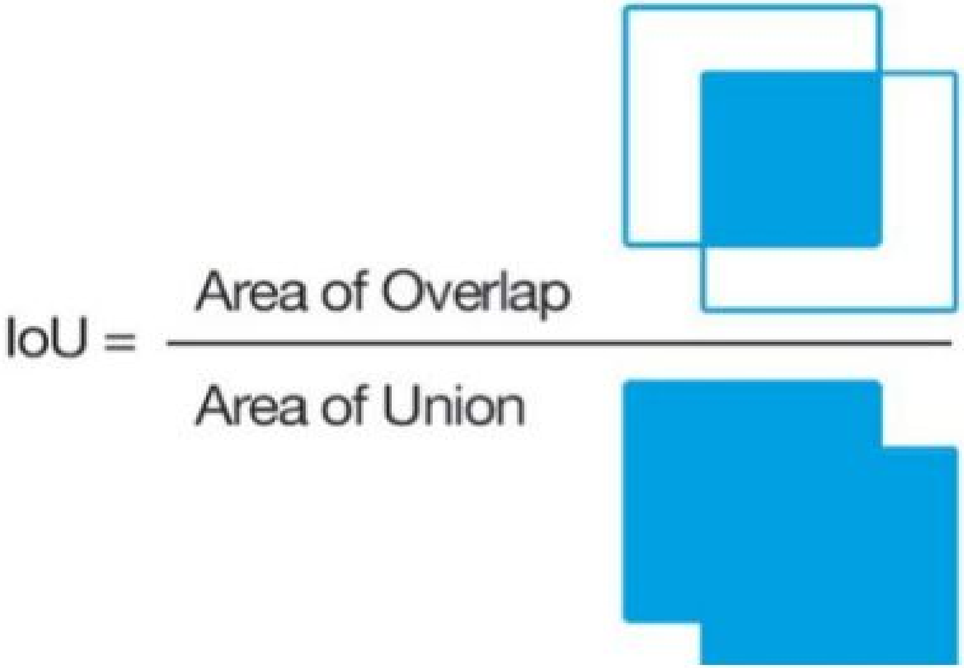
Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).



- 3*3 conv kernel
- 2*2 pooling with stride = 2

Source: https://blog.heuritech.com/2016/02/29/a-brief-report-of-the-heuritech-deep-learning-meetup-5/

Bounding Box Matching Strategy



- Jaccard Overlap **(Intersection over Union)**
- Matching default boxes to ground truth boxes with IoU > threshold(0.5)

Training Objective

- After pairing groundtruth and default boxes, we can write the objective function:

$$L(x,c,l,g) = \frac{1}{N}(L_{conf}(x,c) + \alpha L_{loc}(x,l,g))$$

$x_{ij}^{p} = \{1,0\}$: matching the i-th default box to the j-th ground truth box of category p.

N: matched default boxes.
c: class confidence.
l: predicted bounding box
g: ground truth bounding box

Training Objective

$$L_{loc}(x,l,g) = \sum_{i\in Pos}^{N} \sum_{m\in\{cx,cy,w,h\}} x_{ij}^{k}\,\mathrm{smooth}_{\mathrm{L1}}(l_i^m - \hat{g}_j^m)$$

$$\hat{g}_j^{cx} = (g_j^{cx} - d_i^{cx})/d_i^w \qquad \hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy})/d_i^h$$

$$\hat{g}_j^{w} = \log\left(\frac{g_j^w}{d_i^w}\right) \qquad \hat{g}_j^{h} = \log\left(\frac{g_j^h}{d_i^h}\right)$$

$$L_{conf}(x,c) = -\sum_{i\in Pos}^{N} x_{ij}^{p} log(\hat{c}_i^p) - \sum_{i\in Neg} log(\hat{c}_i^0) \quad \text{where} \quad \hat{c}_i^p = \frac{\exp(c_i^p)}{\sum_p \exp(c_i^p)}$$

Hard Negative Mining

- Instead of using all the negative examples, we sort them using the highest confidence loss for each default box and pick the top ones.

- The ratio between negative examples and positive examples is 3:1.

- This method leads to faster optimization and a more stable training.

Data Augmentation

- Making the model more robust to various input object sizes and outputs:

  1.Original Images.

  2.Sample patch with minimal jaccard scores as 0.1, 0.3, 0.5, 0.7 or 0.9.

  3.Randomly sample a patch.

Experiments

PASCAL VOC2007 test detection results.

Method	data	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Fast [6]	07	66.9	74.5	78.3	69.2	53.2	36.6	77.3	78.2	82.0	40.7	72.7	67.9	79.6	79.2	73.0	69.0	30.1	65.4	70.2	75.8	65.8
Fast [6]	07+12	70.0	77.0	78.1	69.3	59.4	38.3	81.6	78.6	86.7	42.8	78.8	68.9	84.7	82.0	76.6	69.9	31.8	70.1	74.8	80.4	70.4
Faster [2]	07	69.9	70.0	80.6	70.1	57.3	49.9	78.2	80.4	82.0	52.2	75.3	67.2	80.3	79.8	75.0	76.3	39.1	68.3	67.3	81.1	67.6
Faster [2]	07+12	73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6
Faster [2]	07+12+COCO	78.8	84.3	82.0	77.7	68.9	65.7	88.1	88.4	88.9	63.6	86.3	70.8	85.9	87.6	80.1	82.3	53.6	80.4	75.8	86.6	78.9
SSD300	07	68.0	73.4	77.5	64.1	59.0	38.9	75.2	80.8	78.5	46.0	67.8	69.2	76.6	82.1	77.0	72.5	41.2	64.2	69.1	78.0	68.5
SSD300	07+12	74.3	75.5	80.2	72.3	66.3	47.6	83.0	84.2	86.1	54.7	78.3	73.9	84.5	85.3	82.6	76.2	48.6	73.9	76.0	83.4	74.0
SSD300	07+12+COCO	79.6	80.9	86.3	79.0	**76.2**	57.6	87.3	88.2	88.6	60.5	85.4	**76.7**	**87.5**	**89.2**	84.5	81.4	55.0	81.9	**81.5**	85.9	78.9
SSD512	07	71.6	75.1	81.4	69.8	60.8	46.3	82.6	84.7	84.1	48.5	75.0	67.4	82.3	83.9	79.4	76.6	44.9	69.9	69.1	78.1	71.8
SSD512	07+12	76.8	82.4	84.7	78.4	73.8	53.2	86.2	87.5	86.0	57.8	83.1	70.2	84.9	85.2	83.9	79.7	50.3	77.9	73.9	82.5	75.3
SSD512	07+12+COCO	**81.6**	**86.6**	**88.3**	**82.4**	76.0	**66.3**	**88.6**	**88.9**	**89.1**	**65.1**	**88.4**	73.6	86.5	88.9	**85.3**	**84.6**	**59.1**	**85.0**	80.4	**87.4**	**81.2**

mAP: SSD > Faster > Fast
mAP: SSD512 > SSD300   Resolution
More training data is better 07+12+COCO

Table from paper SSD

Experiments

Effects of various design choices and components on SSD performance.

	SSD300				
more data augmentation?		✔	✔	✔	✔
include $\{\frac{1}{2}, 2\}$ box?	✔		✔	✔	✔
include $\{\frac{1}{3}, 3\}$ box?	✔			✔	✔
use atrous?	✔	✔	✔		✔
VOC2007 `test` mAP	65.5	71.6	73.7	74.2	**74.3**

Data Augmentation
More default box shapes is better

Table from Paper SSD

Experiments

Effects of using multiple output layers.

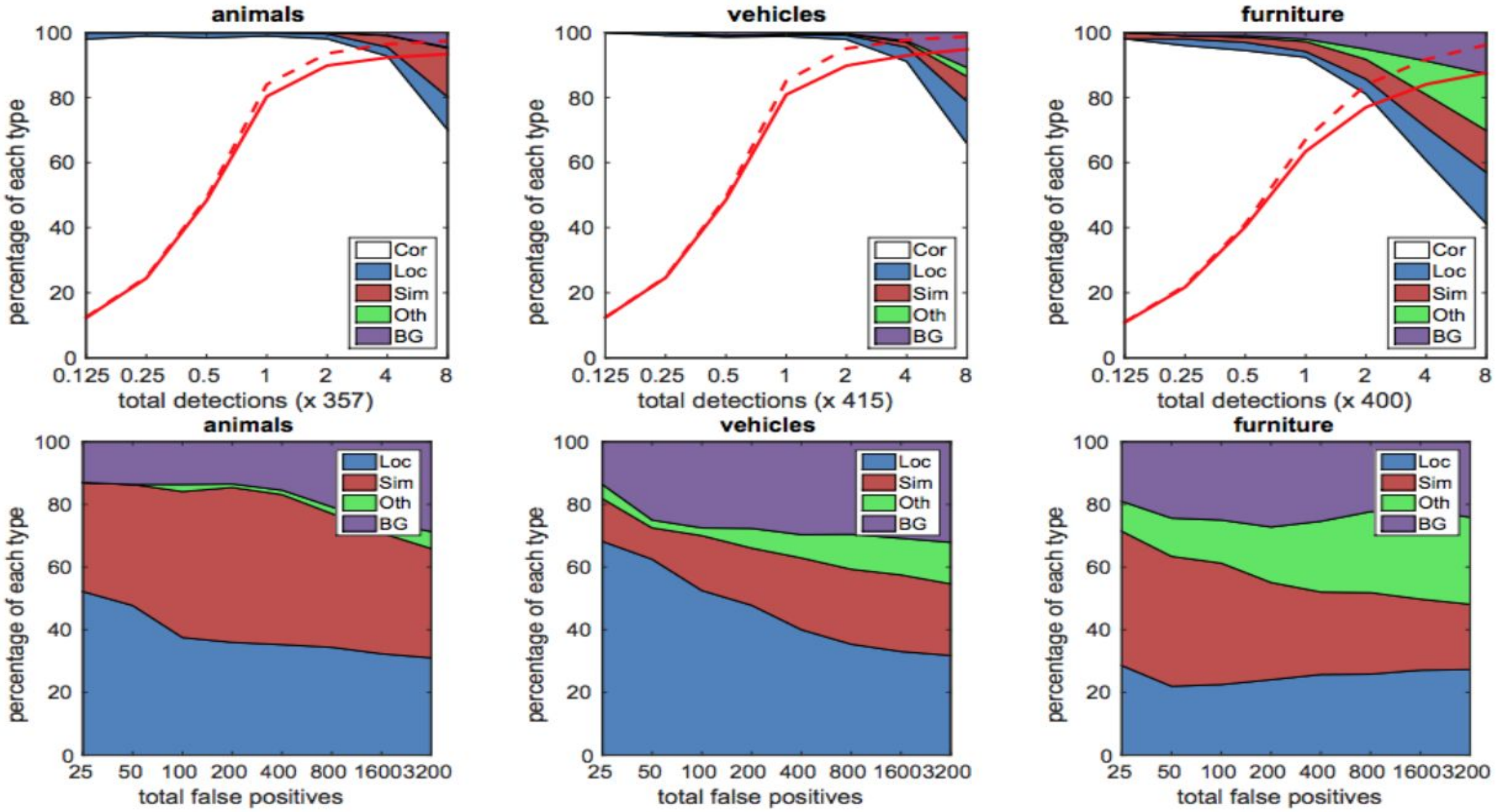
Prediction source layers from:						mAP use boundary boxes?		# Boxes
conv4_3	conv7	conv8_2	conv9_2	conv10_2	conv11_2	Yes	No	
✔	✔	✔	✔	✔	✔	74.3	63.4	8732
✔	✔	✔	✔	✔		**74.6**	63.1	8764
✔	✔	✔	✔			73.8	68.4	8942
✔	✔	✔				70.7	69.2	9864
✔	✔					64.2	64.4	9025
	✔					62.4	64.0	8664

Multiple Output Layers Are Better

Table from paper SSD

Experiments

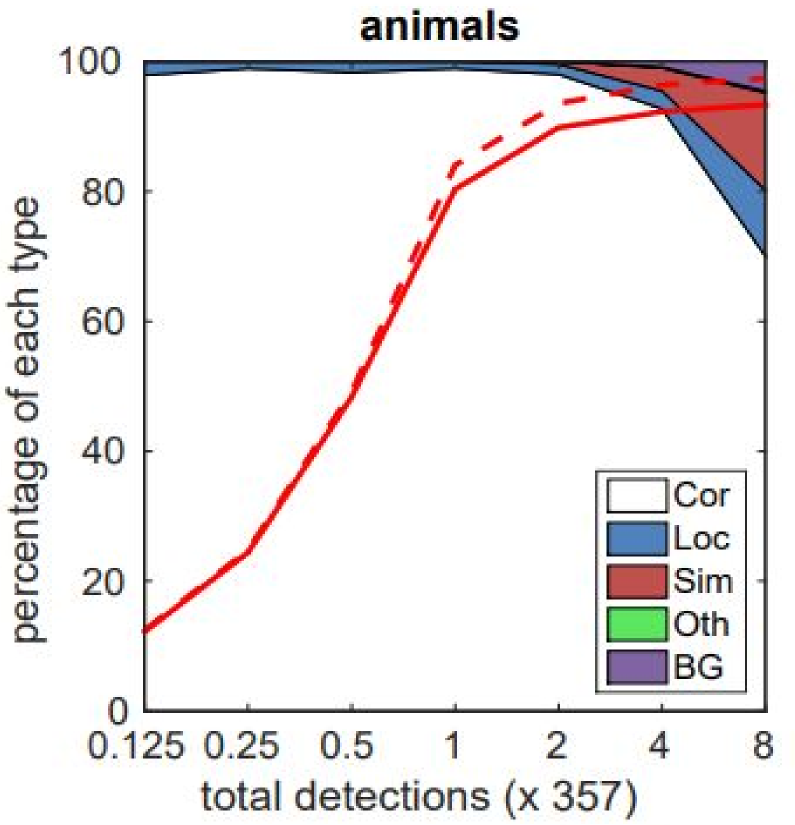
Visulization of performance for SSD512 on animals, vehicles and furniture from VOC2007 test.



Plot from paper SSD

Experiments

Visualization of performance for SSD512 on animals, the cumulative fraction of detections.



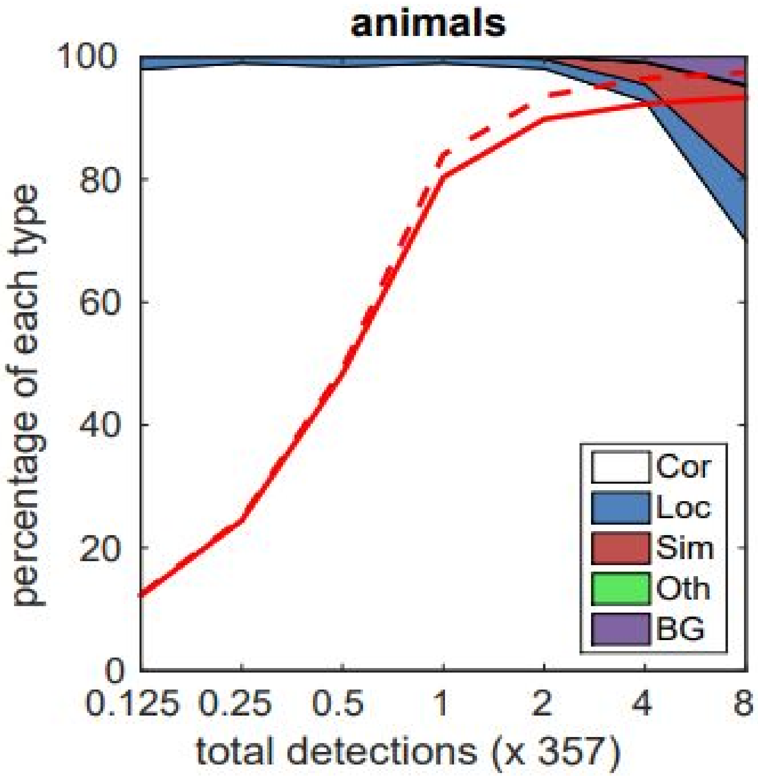
- **Cor:** Correct
- **Loc:** False positive due to poor localization
- **Sim:** Confusion with similar categories
- **Oth:** Confusion with other categories
- **BG:** Confusion with background

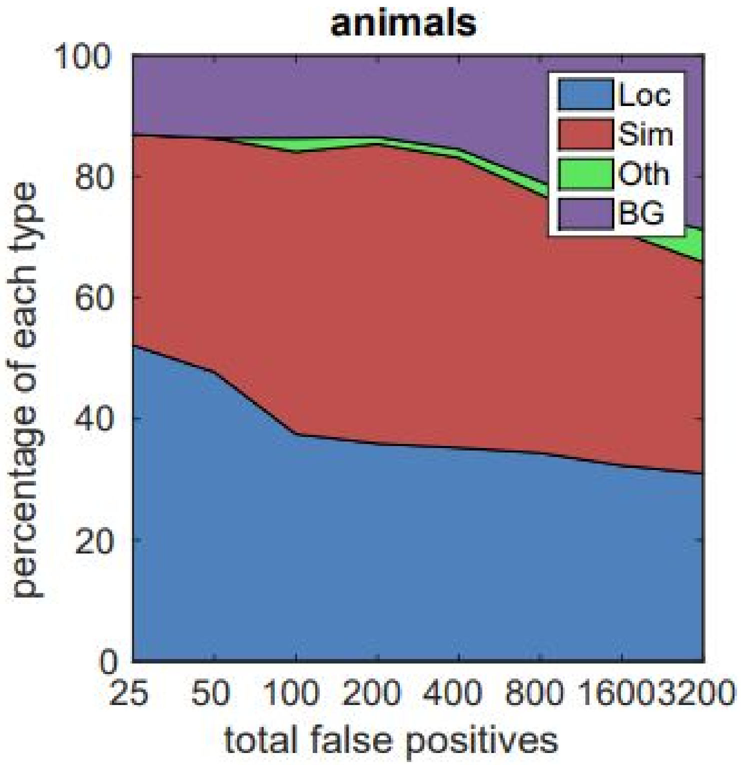
➢ **The change of recall over detections**
- ✓ **The solid red line**: strong criteria (0.5 jaccard overlap)
- ✓ **The dashed red line** : weak criteria

Plot from paper SSD

Experiments

Visualization of performance for SSD512 on animals, the distribution of top-ranked false positive types.





Plot from paper SSD

Experiments

Inference Time(Results on VOC2007 test).

Method	mAP	FPS	batch size	# Boxes	Input resolution
Faster R-CNN (VGG16)	73.2	7	1	~ 6000	$\sim 1000 \times 600$
Fast YOLO	52.7	155	1	98	448×448
YOLO (VGG16)	66.4	21	1	98	448×448
SSD300	74.3	46	1	8732	300×300
SSD512	76.8	19	1	24564	512×512
SSD300	74.3	59	8	8732	300×300
SSD512	76.8	22	8	24564	512×512

Table from paper SSD

Detection Results

Images from paper SSD

Strength

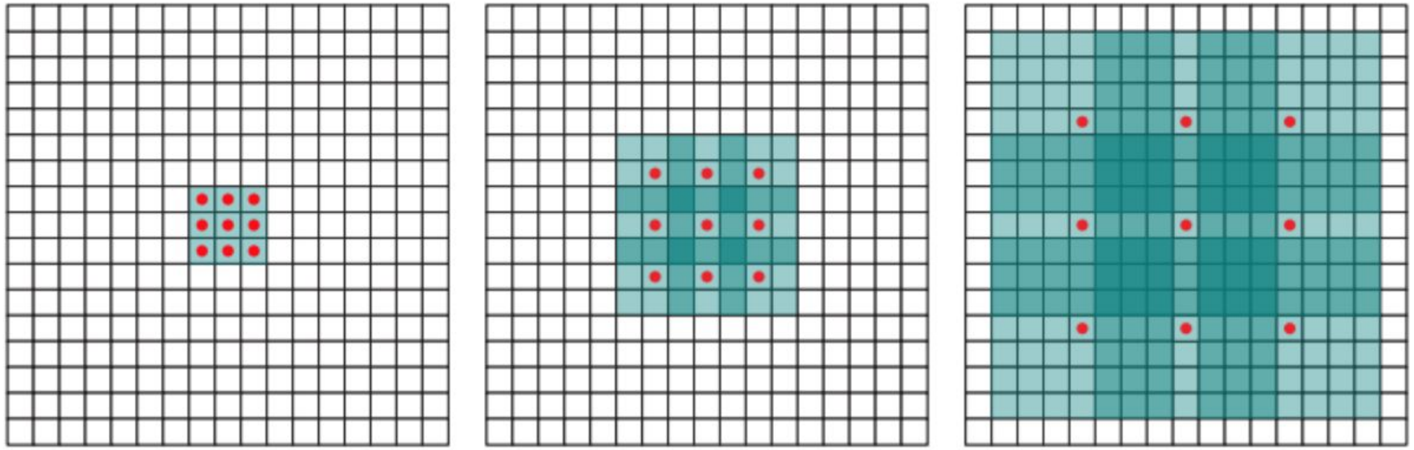
- High Speed
- High Accuracy
- Simple Training(single shot)

Drawbacks

- The classification task for small objects is relatively hard for SSD.

Questions

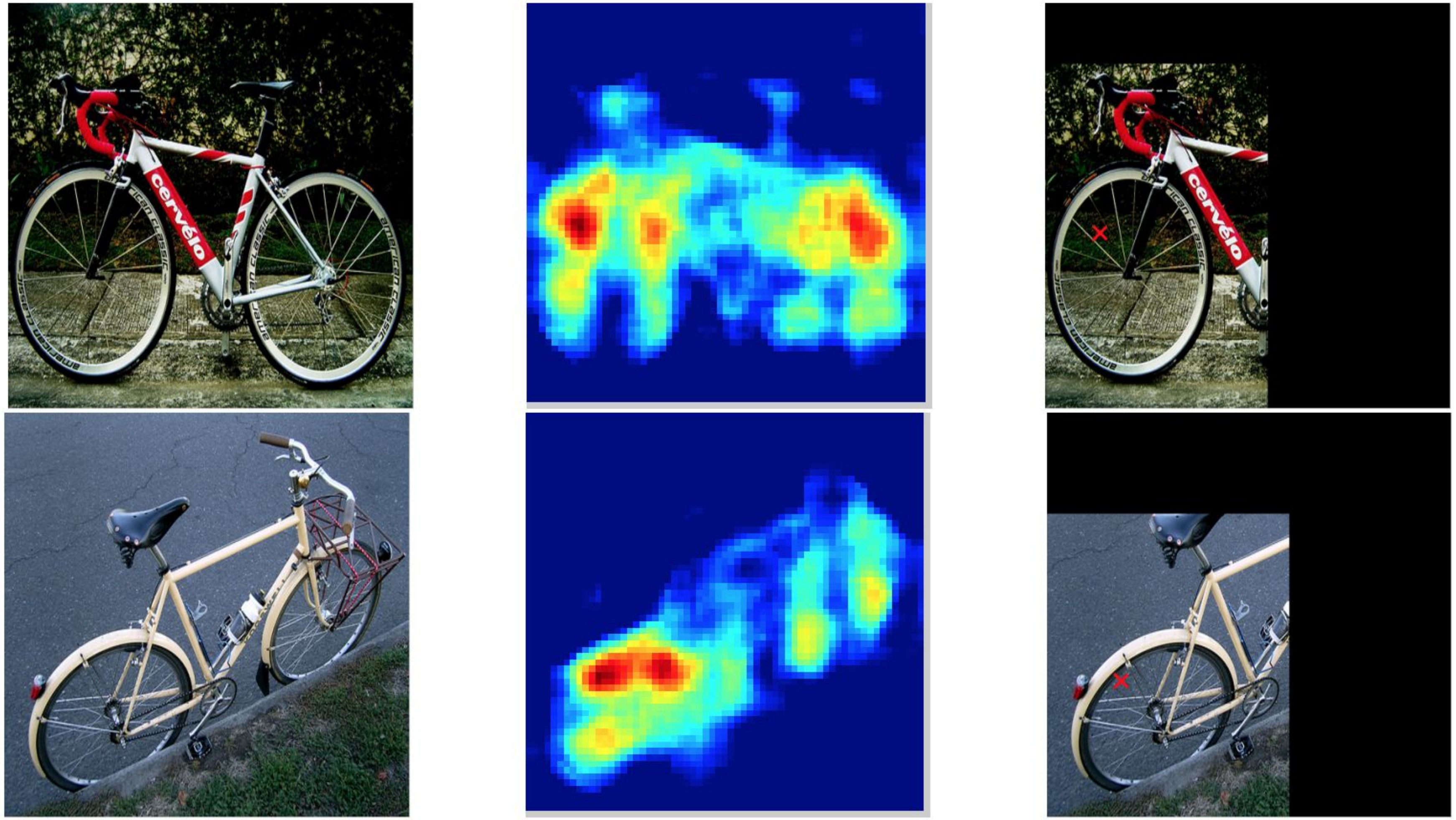
Atrous Algorithm(Dilated Convolution)



(a) (b) (c)

- Figure (a) is a 1-dilated 3x3 convolution filter. In other words, it's a standard 3x3 convolution filter.
- Figure (b) is a 2-dilated 3x3 convolution filter. The red dots are where the weights are and everywhere else is 0. In other words, it's a **5x5 convolution filter with 9 non-zero weights and everywhere else 0**, as mentioned in the question. The receptive field in this case is 7x7 because each unit in the previous output has a receptive field of 3x3. The highlighted portions in blue show the receptive field and **NOT** the convolution filter (you could see it as a convolution filter if you wanted to but it's not helpful).
- Figure (c) is a 4-dilated 3x3 convolution filter. It's a **9x9 convolution filter with 9 non-zeros weights and everywhere else 0**. From (b), we have it that each unit now has a 7x7 receptive field, and hence you can see a 7x7 blue portion around each red dot.

Receptive Field



(a) Original Image (b) Activation map (c) Theoretical RF

L2 Normalization on Conv4

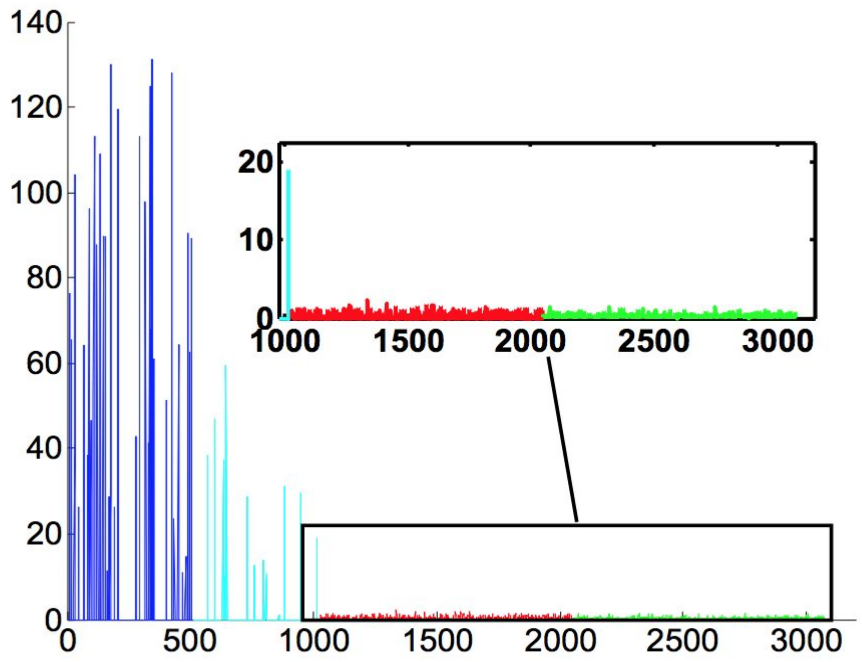


Figure 3: Features from 4 different layers have activations that are of drastically different scales. Each color corresponds to a different layers' feature. While *blue* and *cyan* are on a comparable scale, *red* and *green* features are of a scale 2 orders of magnitude less.

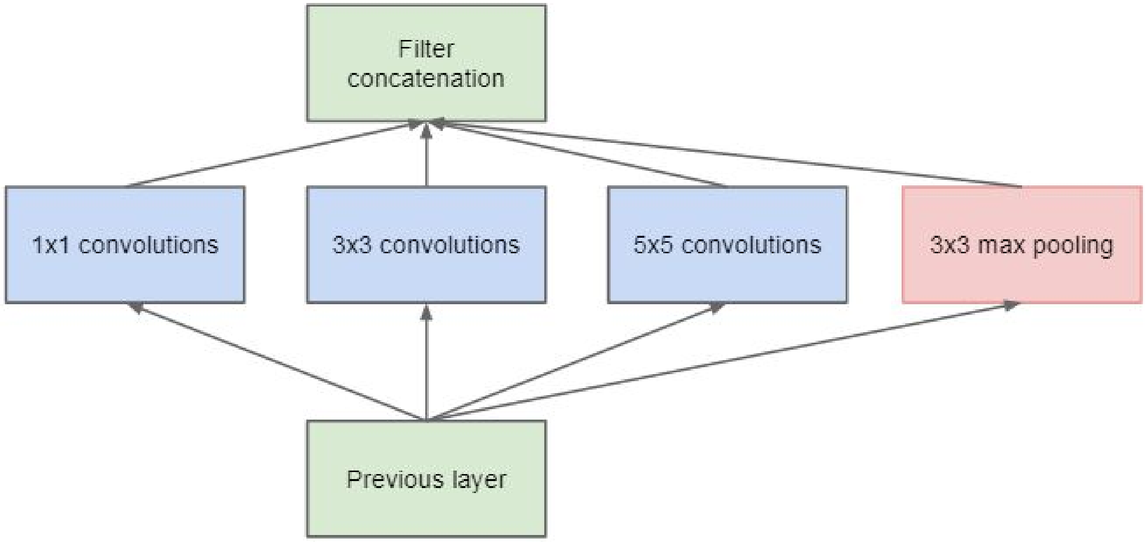
Figure from paper ParseNet

Why Conv6,Conv7?

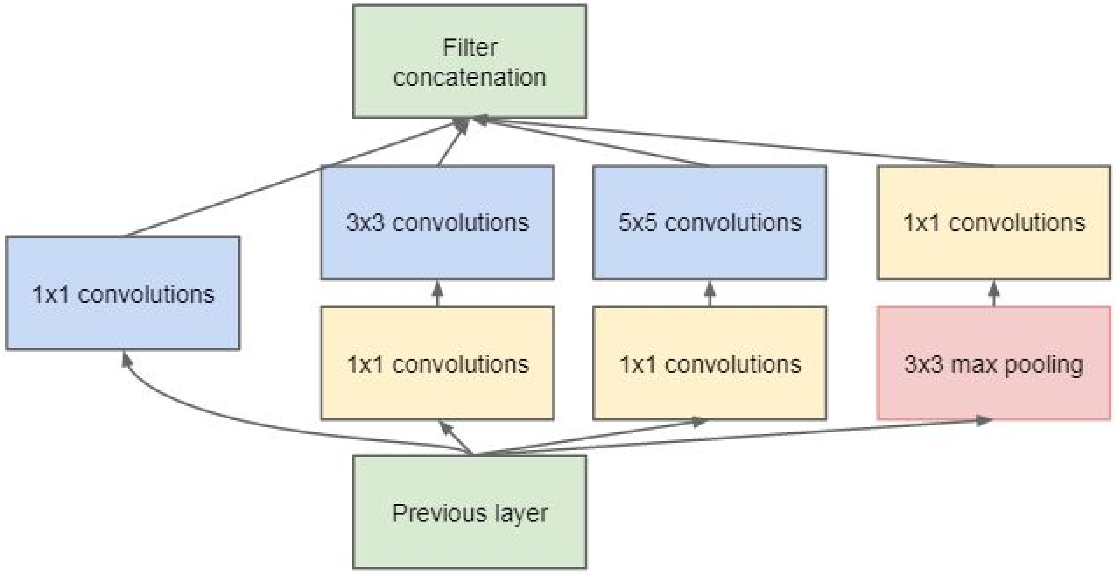
About 1x1 conv kernels

In GoogLeNet architecture, 1x1 convolution is used for two purposes

- To reduce the dimensions inside this “inception module”.
- To add more non-linearity by having ReLU immediately after every 1x1 convolution.



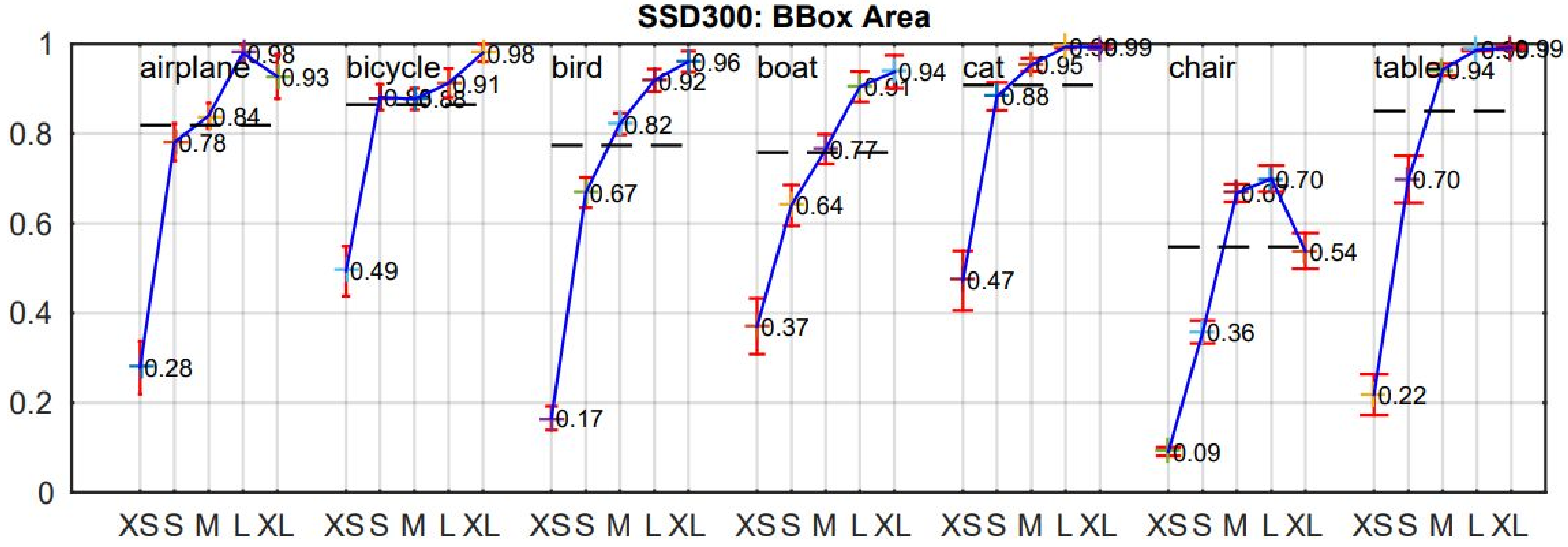
(a) Inception module, naïve version



(b) Inception module with dimension reductions

Experiments

Sensitivity and impact of different object characteristics on VOC2007 test set.



Plot from paper SSD

Object Detection Leaderboard(VOC2012)

Average Precision (AP %)

	mean	aero plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dining table	dog	horse	motor bike	person	potted plant	sheep	sofa	train	tv/ monitor	submission date
R4D_faster_rcnn [?]	**88.6**	94.6	92.3	**91.3**	82.3	79.4	**91.8**	**91.8**	97.4	**76.6**	93.6	**75.3**	97.0	94.6	93.5	92.6	75.1	92.0	80.9	**94.4**	**86.5**	20-Nov-2016
R-FCN, ResNet Ensemble(VOC+COCO) [?]	88.4	94.8	92.9	90.6	**82.4**	**81.8**	89.9	91.7	97.1	76.0	93.4	71.9	96.6	94.3	93.9	**92.8**	**75.7**	91.9	80.8	93.6	86.4	09-Oct-2016
HIK_FRCN [?]	87.9	**95.0**	93.2	**91.3**	80.3	77.7	90.6	89.9	**97.8**	72.8	**93.7**	70.7	**97.2**	**95.4**	**94.0**	91.8	72.7	**92.8**	81.1	94.1	86.2	19-Sep-2016
** Deformable R-FCN, ResNet-101 (VOC+COCO) ** [?]	87.1	94.0	91.7	88.5	79.4	78.0	89.7	90.8	96.9	74.2	93.1	71.3	95.9	94.8	93.2	92.5	71.7	91.8	78.3	93.2	83.3	23-Mar-2017
FasterRcnn-ResNeXt101(COCO+07++12, single model) [?]	86.8	93.9	**93.4**	88.3	80.2	72.6	89.4	89.3	96.8	73.0	91.5	72.3	95.4	94.5	93.8	91.7	70.7	90.6	**81.2**	92.6	83.9	04-May-2017
R-FCN, ResNet (VOC+COCO) [?]	85.0	92.3	89.9	86.7	74.7	75.2	86.7	89.0	95.8	70.2	90.4	66.5	95.0	93.2	92.1	91.1	71.0	89.7	76.0	92.0	83.4	09-Oct-2016
FSSD512 [?]	84.2	92.8	90.0	86.2	75.9	67.7	88.9	89.0	95.0	68.8	90.9	68.7	92.8	92.1	91.4	90.2	63.1	90.1	76.9	91.5	82.7	07-Nov-2017
PVANet+ [?]	84.2	93.5	89.8	84.1	75.6	69.7	88.2	87.9	93.4	70.0	87.7	**75.3**	92.9	90.5	90.9	90.2	67.3	86.4	80.3	92.0	78.8	26-Oct-2016
PFPNet512 VGG16 07++12+COCO [?]	83.8	93.0	89.9	85.1	75.8	66.4	88.4	88.3	94.0	67.9	89.5	69.7	92.0	91.8	91.6	88.7	61.1	89.1	78.4	90.5	84.3	18-Oct-2017
BlitzNet512 [?]	83.8	93.1	89.4	84.7	75.5	65.0	86.6	87.4	94.5	69.9	88.8	71.7	92.5	91.6	91.1	88.9	61.2	90.4	79.2	91.8	83.0	19-Jul-2017
Faster RCNN, ResNet (VOC+COCO) [?]	83.8	92.1	88.4	84.8	75.9	71.4	86.3	87.8	94.2	66.8	89.4	69.2	93.9	91.9	90.9	89.6	67.9	88.2	76.8	90.3	80.0	10-Dec-2015
PVANet+ (compressed) [?]	83.7	92.8	88.9	83.4	74.7	68.7	88.2	87.8	93.5	69.5	87.3	74.3	93.1	89.5	89.9	90.2	66.8	86.4	79.8	91.9	78.2	18-Nov-2016
Cascaded_CrystalNet [?]	83.6	92.6	89.5	83.5	74.7	69.7	87.5	87.6	92.9	70.0	86.9	75.0	91.6	89.5	90.6	90.2	67.2	85.2	80.0	91.4	76.9	23-Dec-2017
DOH_512 (single VGG16, COCO+VOC07++12) [?]	83.4	93.0	89.8	84.5	74.3	63.2	89.3	88.2	94.2	68.0	88.0	69.1	92.3	91.4	90.2	89.0	62.6	89.2	76.7	90.8	83.2	07-Nov-2017
ICT_360_ISD [?]	82.6	90.7	89.4	87.0	75.8	70.1	86.0	86.5	96.2	65.3	86.8	62.1	94.6	90.6	90.5	89.7	63.5	87.3	72.7	90.7	77.1	18-Nov-2016
Rank of experts (VOC07++12) [?]	82.2	90.4	87.4	85.3	72.9	70.8	84.5	87.2	95.6	64.6	87.1	65.4	94.3	89.7	89.5	89.2	66.0	85.1	72.5	89.6	76.6	15-Nov-2017
SSD512 VGG16 07++12+COCO [?]	82.2	91.4	88.6	82.6	71.4	63.1	87.4	88.1	93.9	66.9	86.6	66.3	92.0	91.7	90.8	88.5	60.9	87.0	75.4	90.2	80.4	10-Oct-2016
FSSD300 [?]	82.0	92.2	89.2	81.8	72.3	59.7	87.4	84.4	93.5	66.8	87.7	70.4	92.1	90.9	89.6	87.7	56.9	86.8	79.0	90.7	81.3	10-Nov-2017
RUN_3WAY_300, VGG16, 07++12+COCO [?]	81.7	91.5	88.6	80.3	71.2	59.6	86.4	84.2	94.1	66.6	86.5	70.4	92.1	90.5	89.6	87.5	57.7	86.7	79.6	90.4	80.2	13-Oct-2017
YOLOv2 (VOC + COCO) [?]	81.5	90.0	88.6	82.2	71.7	65.5	85.5	84.2	92.9	67.2	87.6	70.0	91.2	90.5	90.0	88.6	62.5	83.8	70.7	88.8	79.4	21-Oct-2017