Faster R-CNN:

Towards Real-Time Object Detection with Region Proposal Networks

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun

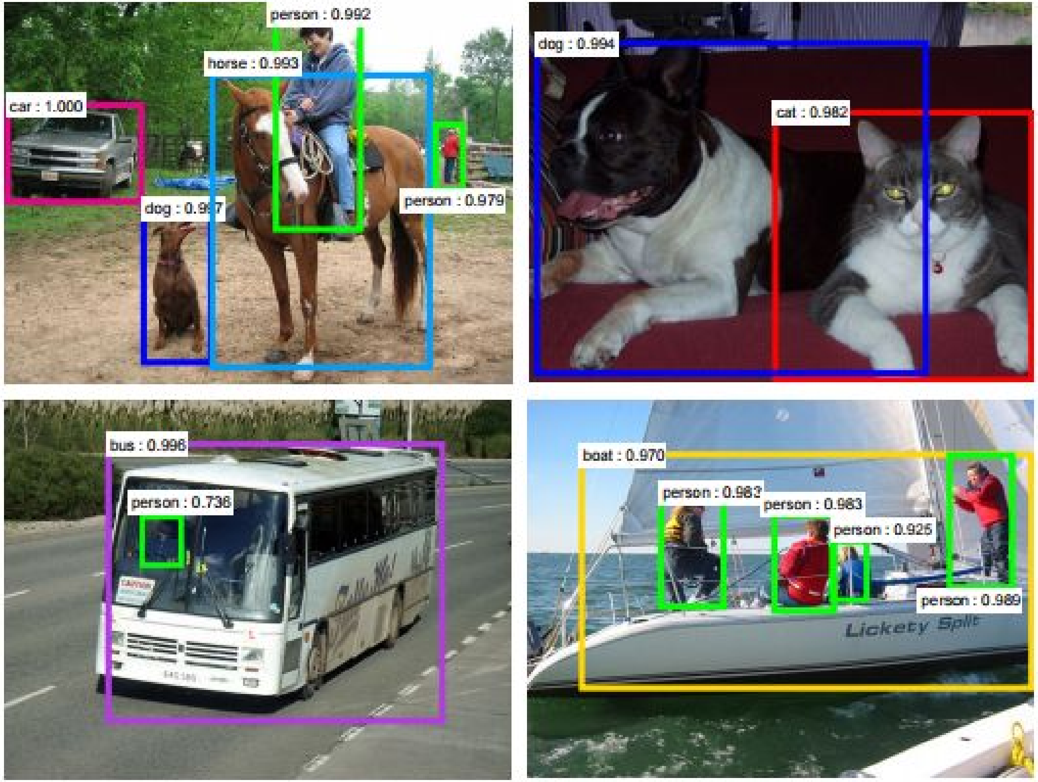
[paper@NIPS15][arXiv][python][matlab][slides by R. Girshick]

Image Processing Group
Signal Theory and Communications Department
Universitat Politècnica de Catalunya. BARCELONATECH

Slides by Amaia Salvador [GDoc]
Computer Vision Reading Group (01/03/2016)

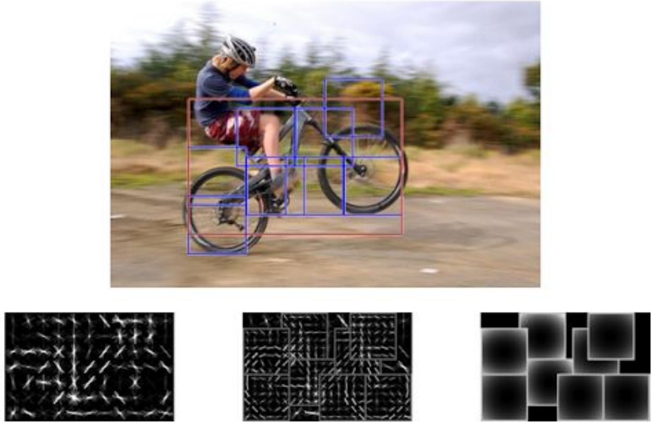
1. Introduction

Object Detection



Object Detection: Previously...

Hand-crafted features + Sliding Window



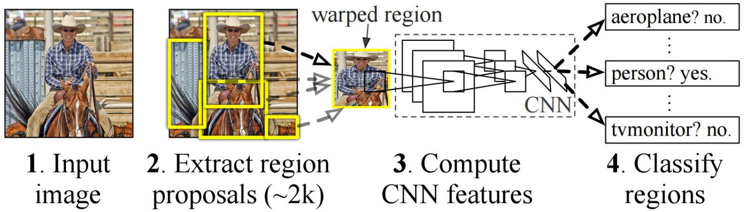
DPM

DPM. P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan. Object Detection with Discriminatively Trained Part Based Models. In IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 9, Sep. 2010

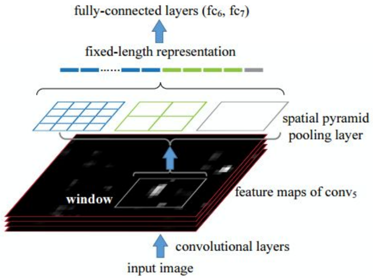
Object Detection: Previously...

CNN features + Object Proposals

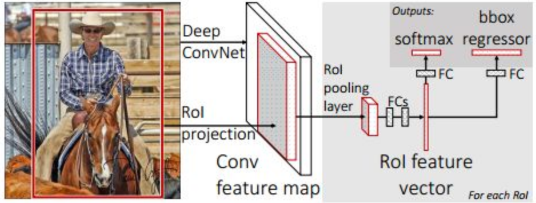
R-CNN



SPPnet

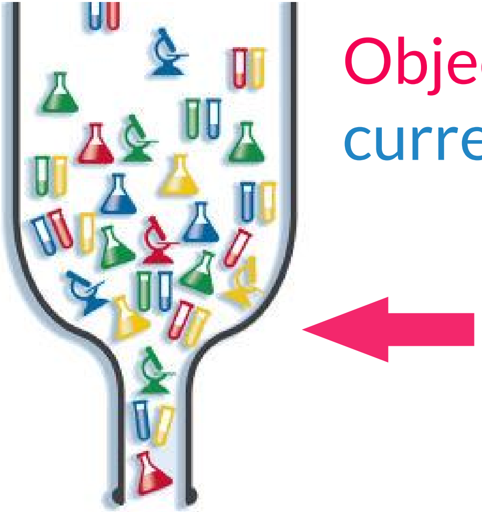


Fast R-CNN



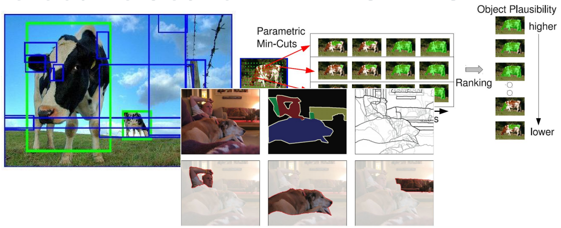
R-CNN. Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014, June). Rich feature hierarchies for accurate object detection and semantic segmentation. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on(pp. 580-587). IEEE.
SPPnet. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 37(9), 1904-1916.
Fast R-CNN. Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1440-1448).

Object Detection: Limitations



Object Proposal computation is the bottleneck in current state of the art object detection systems

Selective Search CPMC



MCG

Selective Search. Van de Sande, K. E., Uijlings, J. R., Gevers, T., & Smeulders, A. W. (2011, November). Segmentation as selective search for object recognition. InComputer Vision (ICCV), 2011 IEEE International Conference on (pp. 1879-1886). IEEE.
CPMC. Carreira, J., & Sminchisescu, C. (2010, June). Constrained parametric min-cuts for automatic object segmentation. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on (pp. 3241-3248). IEEE.
MCG. Arbeláez, P., Pont-Tuset, J., Barron, J., Marques, F., & Malik, J. (2014). Multiscale combinatorial grouping. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 328-335).

Faster R-CNN: Motivation

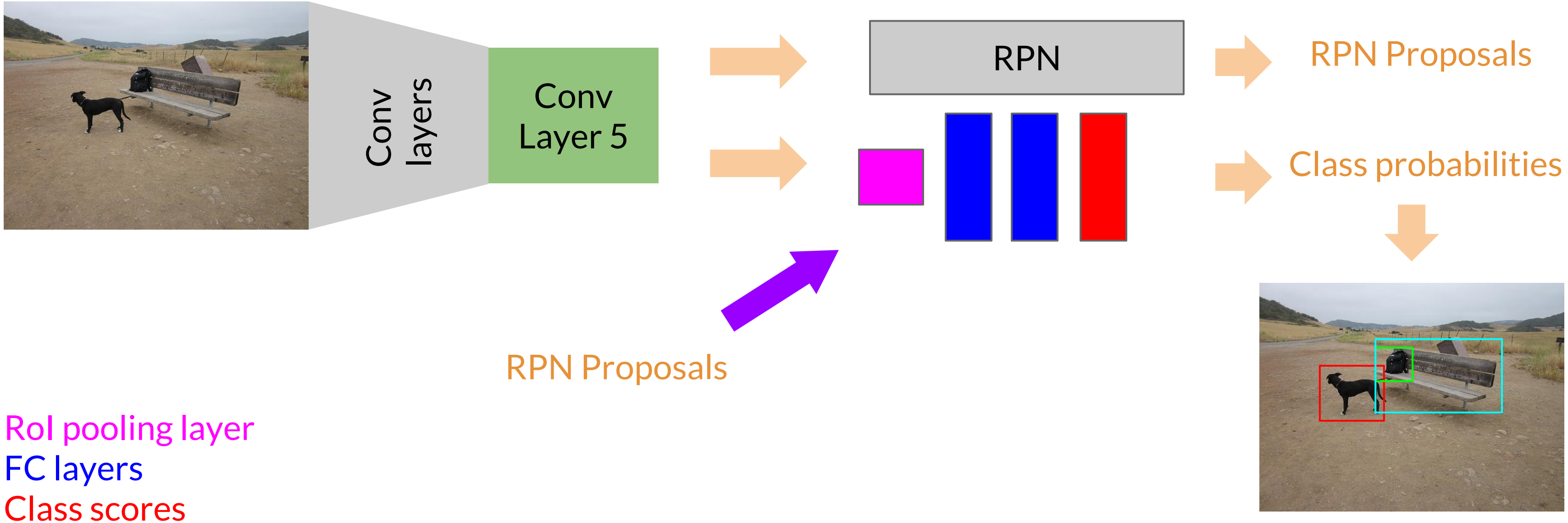
Replace the usage of external Object Proposals with a Region Proposal Network (RPN).

Selective Search CPMC

MCG

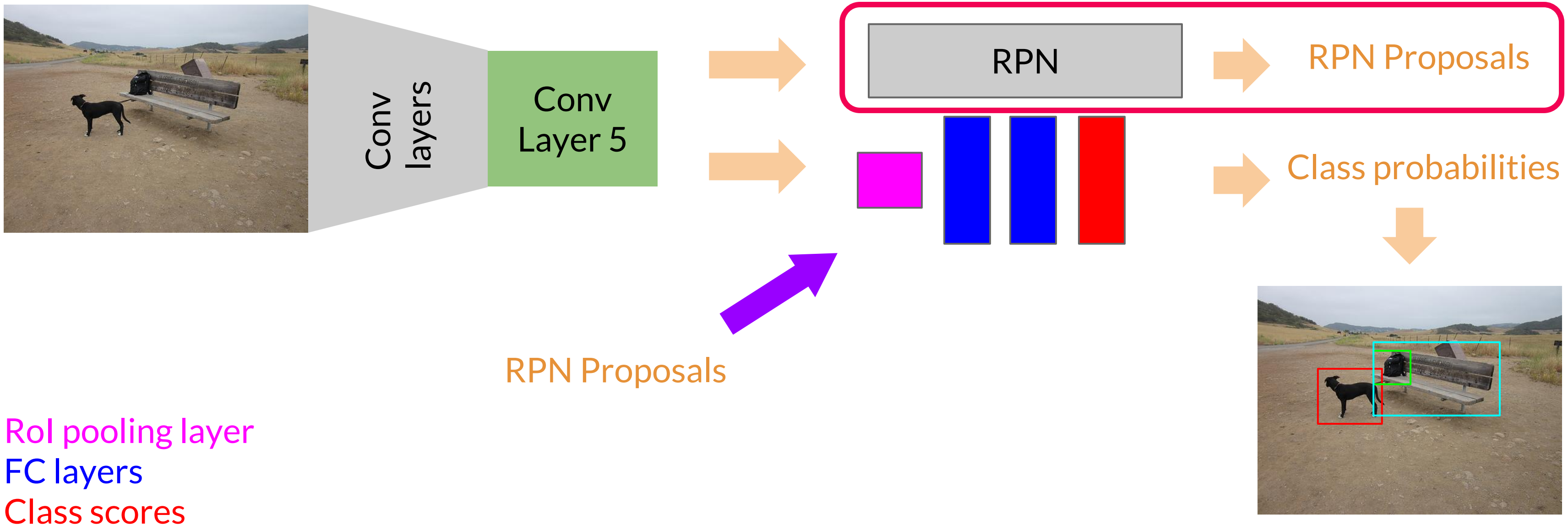
2. Methodology

Faster R-CNN: Overview



RoI pooling layer
FC layers
Class scores

Faster R-CNN: Overview

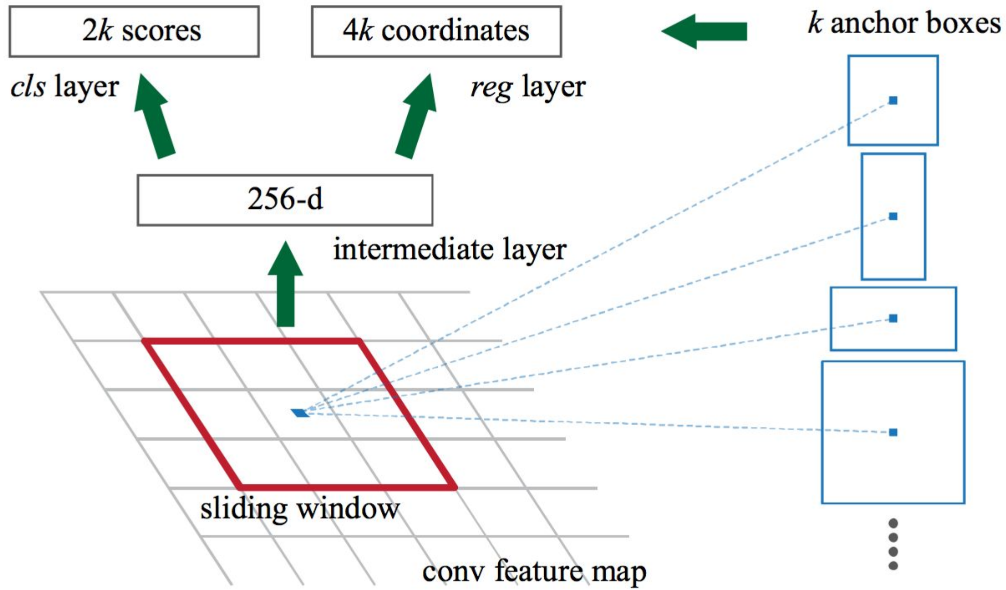


RoI pooling layer
FC layers
Class scores

Region Proposal Network (RPN)

Bounding Box Regression

Objectness scores



In practice, k = 9 (3 different scales and 3 aspect ratios)

RPN: Loss Function

i = anchor index in minibatch

Ground truth objectness label

Log loss

Smooth
L1 loss

$$L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*).$$

True box coordinates

Coordinates of the predicted bounding box for anchor *i*

Predicted probability of being an object for anchor *i*

In practice λ= 10, so that both terms are roughly equally balanced

N_{cls} = Number of anchors in minibatch (~ 256)
N_{reg} = Number of anchor locations (~ 2400)

RPN: Positive/Negative Samples

An anchor is labeled as positive if:

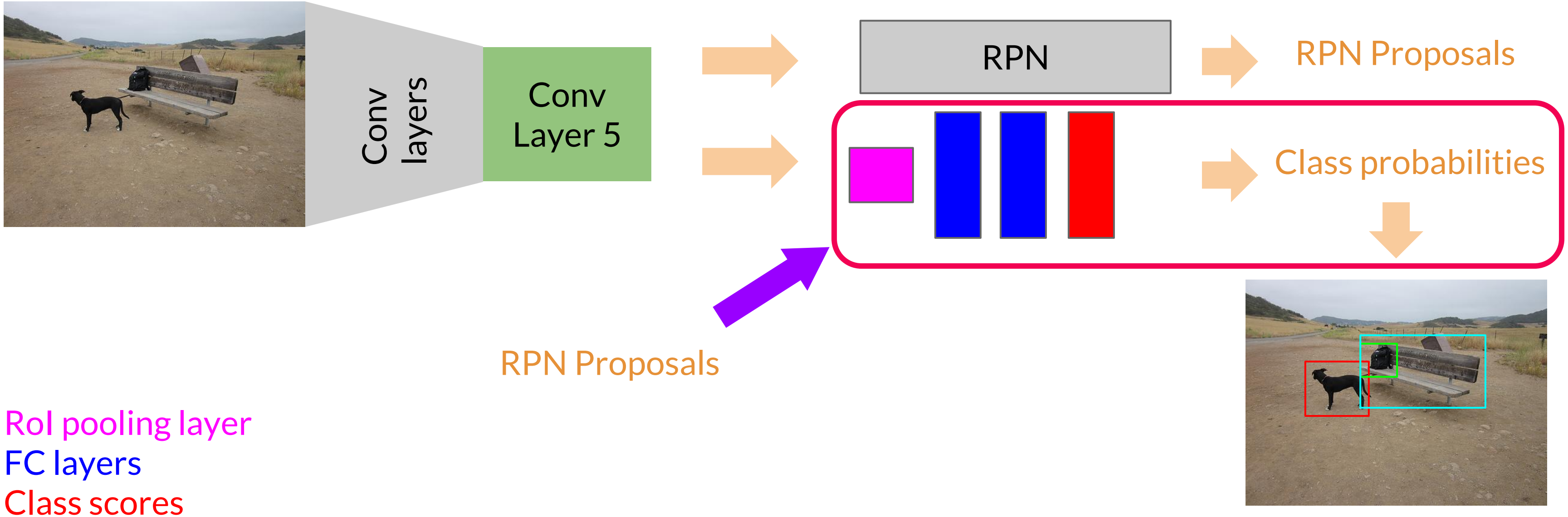
(a) the anchor is the one with highest IoU overlap with a ground-truth box

(b) the anchor has an IoU overlap with a ground-truth box higher than 0.7

Negative labels are assigned to anchors with IoU lower than 0.3 for all ground-truth boxes.

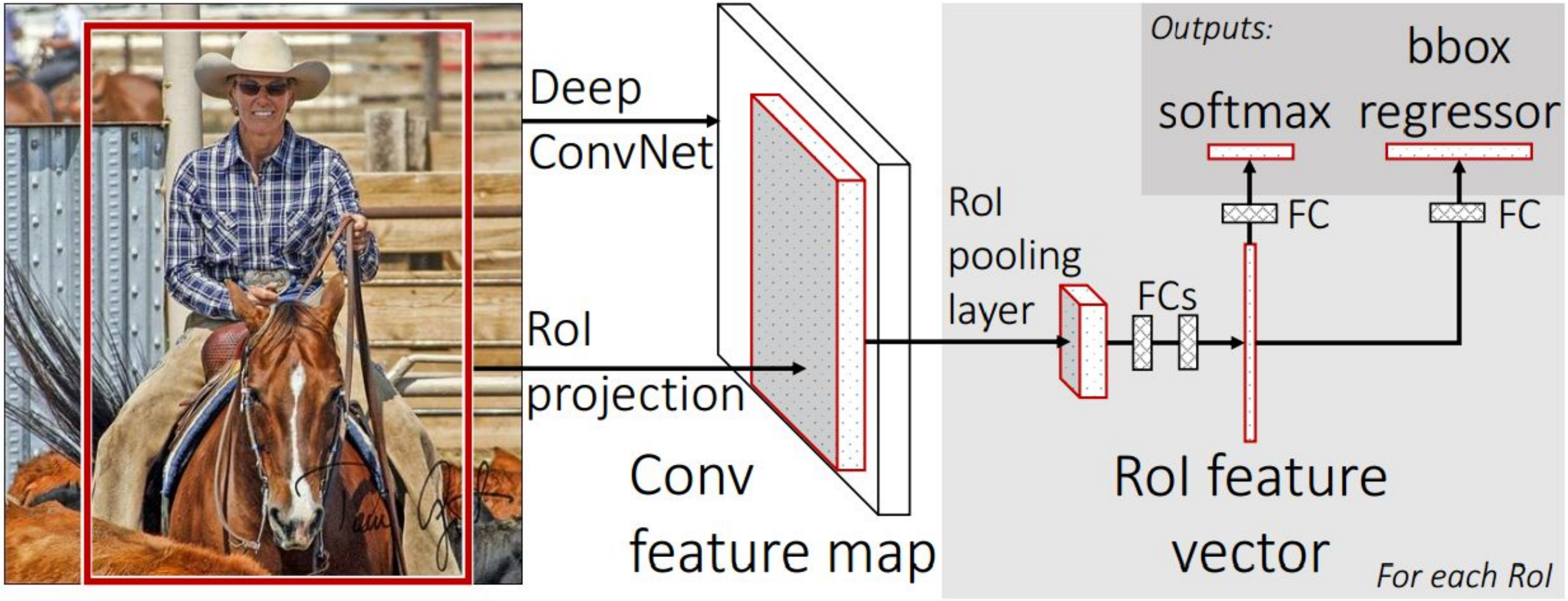
50%/50% ratio of positive/negative anchors in a minibatch.

Faster R-CNN: Overview



RoI pooling layer
FC layers
Class scores

Object Detection Network



Fast R-CNN

Object Detection Network: Loss

*From Fast R-CNN

True box coordinates

Predicted box coordinates

$$L(p, u, t^u, v) = L_{\text{cls}}(p, u) + \lambda[u \geq 1]L_{\text{loc}}(t^u, v),$$

Log loss

True class scores

Smooth
L1 loss

Predicted class scores

Fast R-CNN: Positive/Negative Samples

*From Fast R-CNN

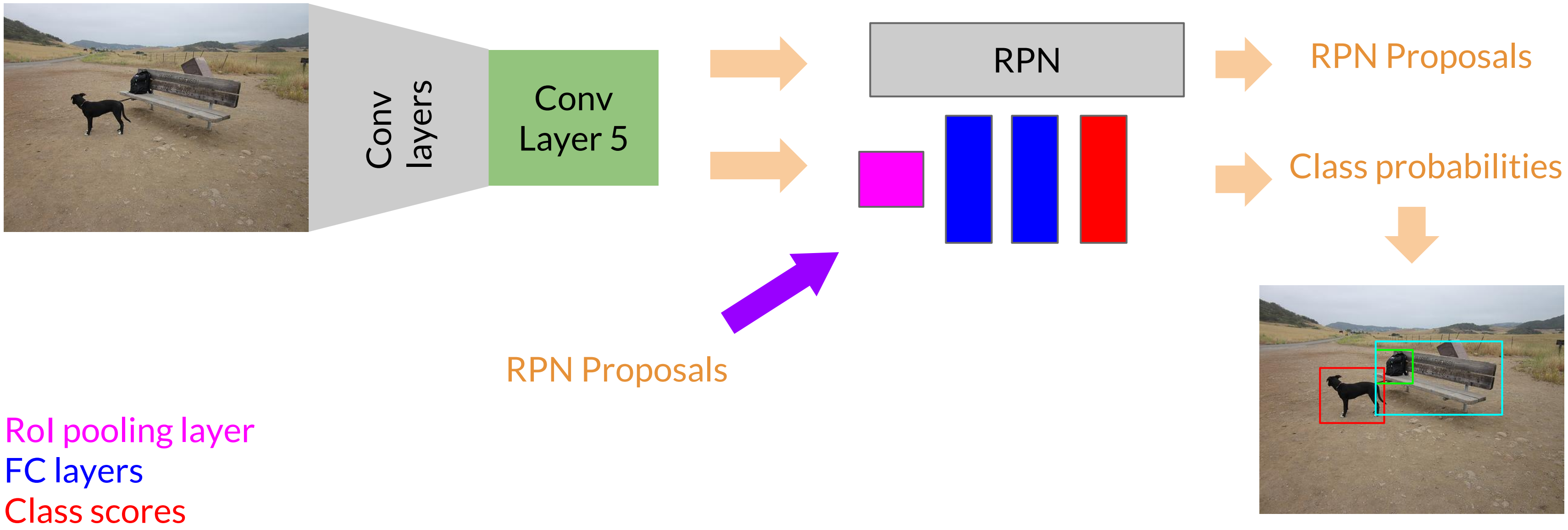
Positive samples are defined as those whose IoU overlap with a ground-truth bounding box is > 0.5.

Negative examples are sampled from those that have a maximum IoU overlap with ground truth in the interval [0.1, 0.5).

25%/75% ratio for positive/negative samples in a minibatch.

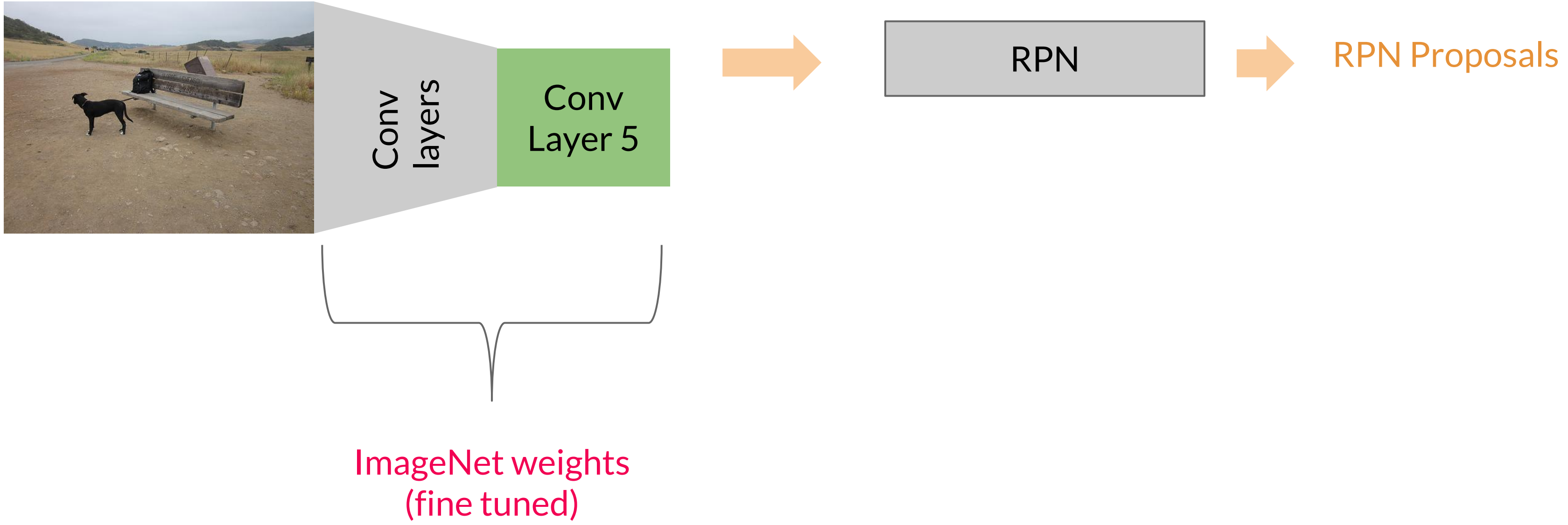
Faster R-CNN: Training

4-step training to share features for RPN and Fast R-CNN



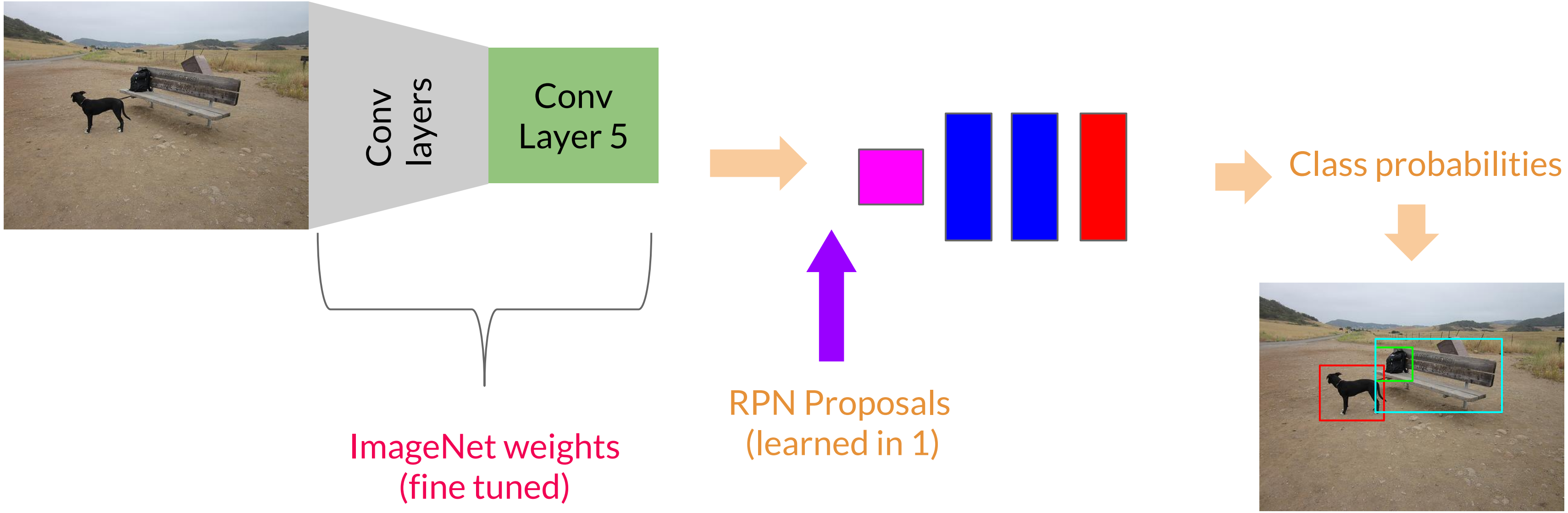
Faster R-CNN: 4-step training

Step 1: Train RPN initialized with an ImageNet pre-trained model.



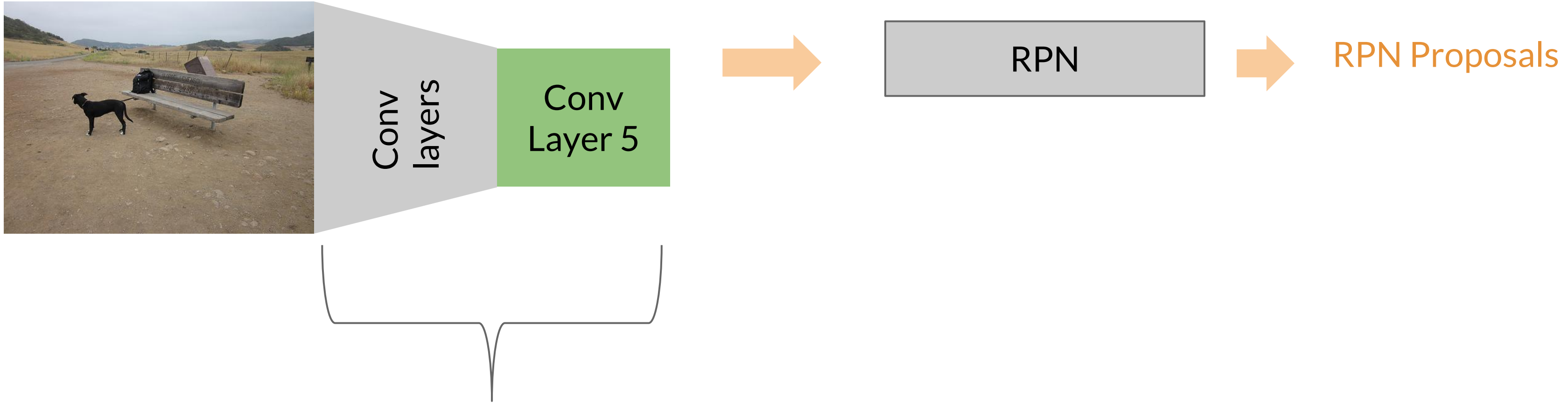
Faster R-CNN: 4-step training

Step 2: Train Fast R-CNN with learned RPN proposals.



Faster R-CNN: 4-step training

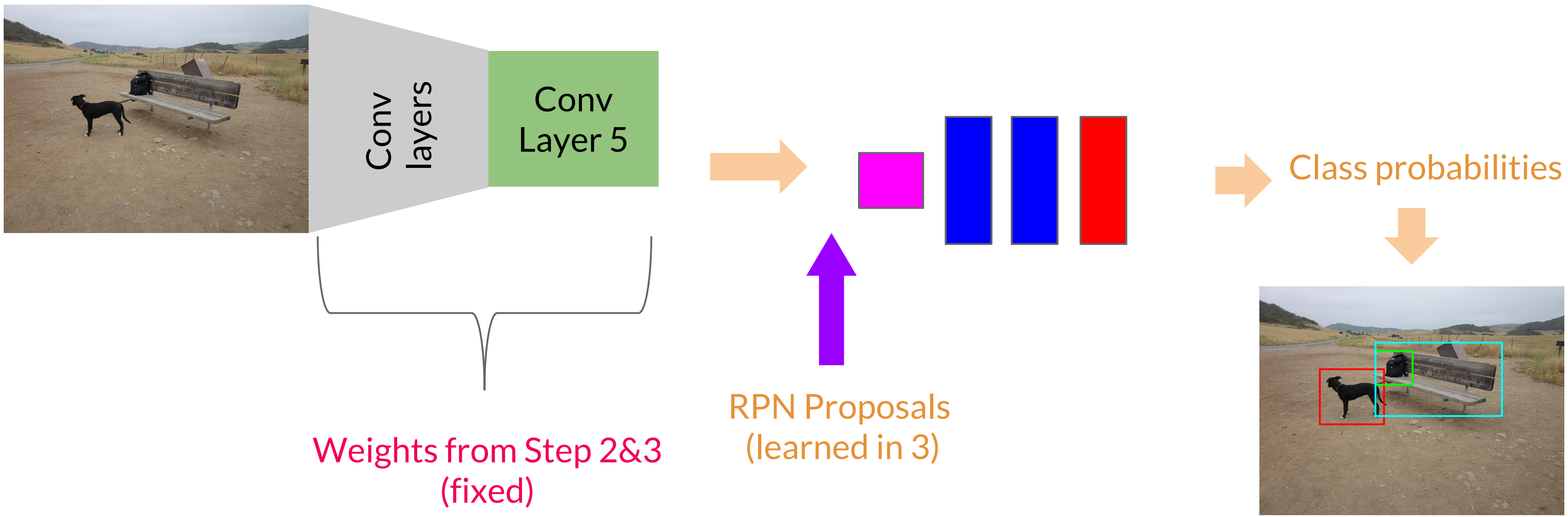
Step 3: The model trained in 2 is used to initialize RPN and train again.



Weights from Step 2
(fixed)

Faster R-CNN: 4-step training

Step 4: Fine tune FC layers of Fast R-CNN using same shared convolutional layers as in 3.



3. Experiments

Experiments: CNN Architectures

VGG-16: Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

ZF: Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In Computer vision–ECCV 2014 (pp. 818-833). Springer International Publishing.

Experiments: Datasets

Visual Object Classes Challenge 2012 (VOC2012)

PASCAL2

Pattern Analysis, Statistical Modelling and Computational Learning

Microsoft COCO

Common Objects in Context

Experiments I: VOC 2007 & ZF

train-time region proposals		test-time region proposals		
method	# boxes	method	# proposals	mAP (%)
SS	2k	SS	2k	58.7
EB	2k	EB	2k	58.6
RPN+ZF, shared	2k	RPN+ZF, shared	300	**59.9**

Comparison between Fast R-CNN trained with external object proposals
(SS: Selective Search, EB: EdgeBoxes) with Faster R-CNN

Experiments I: VOC 2007 & ZF

train-time region proposals		test-time region proposals		
method	# boxes	method	# proposals	mAP (%)
SS	2k	SS	2k	58.7
EB	2k	EB	2k	58.6
RPN+ZF, shared	2k	RPN+ZF, shared	300	**59.9**
ablation experiments follow below				
RPN+ZF, unshared	2k	RPN+ZF, unshared	300	58.7
SS	2k	RPN+ZF	100	55.1
SS	2k	RPN+ZF	300	56.8
SS	2k	RPN+ZF	1k	56.3
SS	2k	RPN+ZF (no NMS)	6k	55.2
SS	2k	RPN+ZF (no *cls*)	100	44.6
SS	2k	RPN+ZF (no *cls*)	300	51.4
SS	2k	RPN+ZF (no *cls*)	1k	55.8
SS	2k	RPN+ZF (no *reg*)	300	52.1
SS	2k	RPN+ZF (no *reg*)	1k	51.3
SS	2k	RPN+VGG	300	59.2

Experiments I: VOC 2007 & ZF

train-time region proposals		test-time region proposals		
method	# boxes	method	# proposals	mAP (%)
SS	2k	SS	2k	58.7
EB	2k	EB	2k	58.6
RPN+ZF, shared	2k	RPN+ZF, shared	300	**59.9**
ablation experiments follow below				
RPN+ZF, unshared	2k	RPN+ZF, unshared	300	58.7
SS	2k	RPN+ZF	100	55.1
SS	2k	RPN+ZF	300	56.8
SS	2k	RPN+ZF	1k	56.3
SS	2k	RPN+ZF (no NMS)	6k	55.2
SS	2k	RPN+ZF (no *cls*)	100	44.6
SS	2k	RPN+ZF (no *cls*)	300	51.4
SS	2k	RPN+ZF (no *cls*)	1k	55.8
SS	2k	RPN+ZF (no *reg*)	300	52.1
SS	2k	RPN+ZF (no *reg*)	1k	51.3
SS	2k	RPN+VGG	300	59.2

Experiments I: VOC 2007 & ZF

train-time region proposals method	# boxes	**test-time region proposals** method	# proposals	mAP (%)
SS	2k	SS	2k	58.7
EB	2k	EB	2k	58.6
RPN+ZF, shared	2k	RPN+ZF, shared	300	**59.9**
ablation experiments follow below				
RPN+ZF, unshared	2k	RPN+ZF, unshared	300	58.7
SS	2k	RPN+ZF	100	55.1
SS	2k	RPN+ZF	300	56.8
SS	2k	RPN+ZF	1k	56.3
SS	2k	RPN+ZF (no NMS)	6k	55.2
SS	2k	RPN+ZF (no *cls*)	100	44.6
SS	2k	RPN+ZF (no *cls*)	300	51.4
SS	2k	RPN+ZF (no *cls*)	1k	55.8
SS	2k	RPN+ZF (no *reg*)	300	52.1
SS	2k	RPN+ZF (no *reg*)	1k	51.3
SS	2k	RPN+VGG	300	59.2

Experiments II

Detection Accuracy

method	# proposals	data	mAP (%)
SS	2k	07	$66.9^{\dagger}$
SS	2k	07+12	70.0
RPN+VGG, unshared	300	07	68.5
RPN+VGG, shared	300	07	69.9
RPN+VGG, shared	300	07+12	**73.2**

Timing (ms)

model	system	conv	proposal	region-wise	total	rate
VGG	SS + Fast R-CNN	146	1510	174	1830	0.5 fps
VGG	RPN + Fast R-CNN	141	**10**	47	**198**	**5 fps**
ZF	RPN + Fast R-CNN	31	**3**	25	**59**	**17 fps**

Experiments III

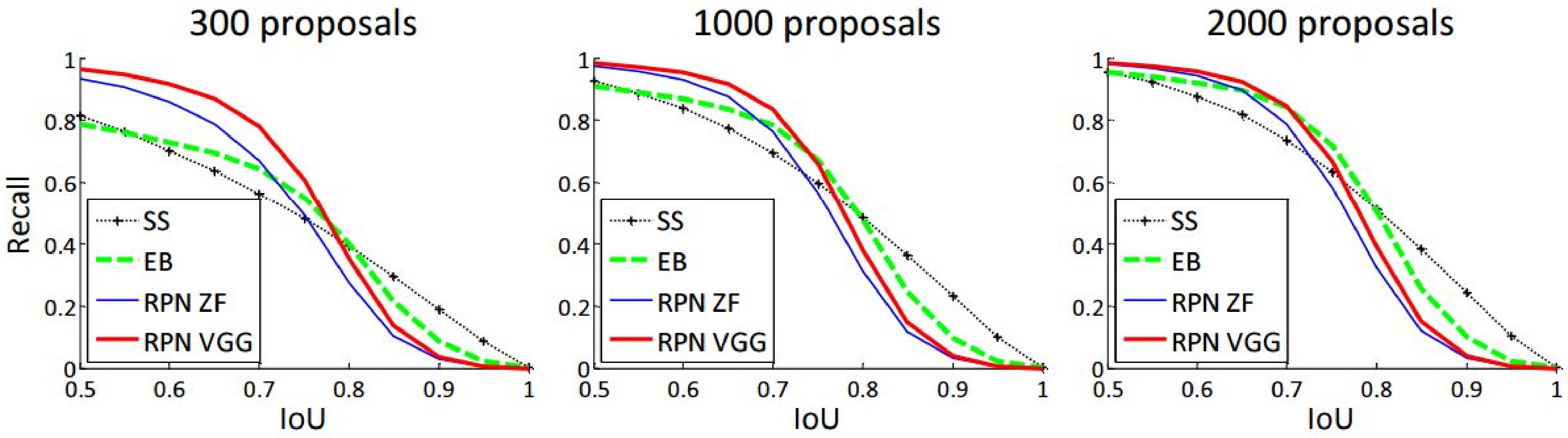


Figure 2: Recall *vs.* IoU overlap ratio on the PASCAL VOC 2007 test set.

Experiments IV

One-Stage Detection:

1) Directly Refine and Classify Sliding Window locations

Two-Stage Proposal + Detection:

1) Learn Object Proposals
2) Refine and classify Object Proposals

Table 5: **One-Stage Detection *vs.* Two-Stage Proposal + Detection**. Detection results are on the PASCAL VOC 2007 test set using the ZF model and Fast R-CNN. RPN uses unshared features.

	regions		detector	mAP (%)
Two-Stage	RPN + ZF, unshared	300	Fast R-CNN + ZF, 1 scale	58.7
One-Stage	dense, 3 scales, 3 asp. ratios	20k	Fast R-CNN + ZF, 1 scale	53.8
One-Stage	dense, 3 scales, 3 asp. ratios	20k	Fast R-CNN + ZF, 5 scales	53.9

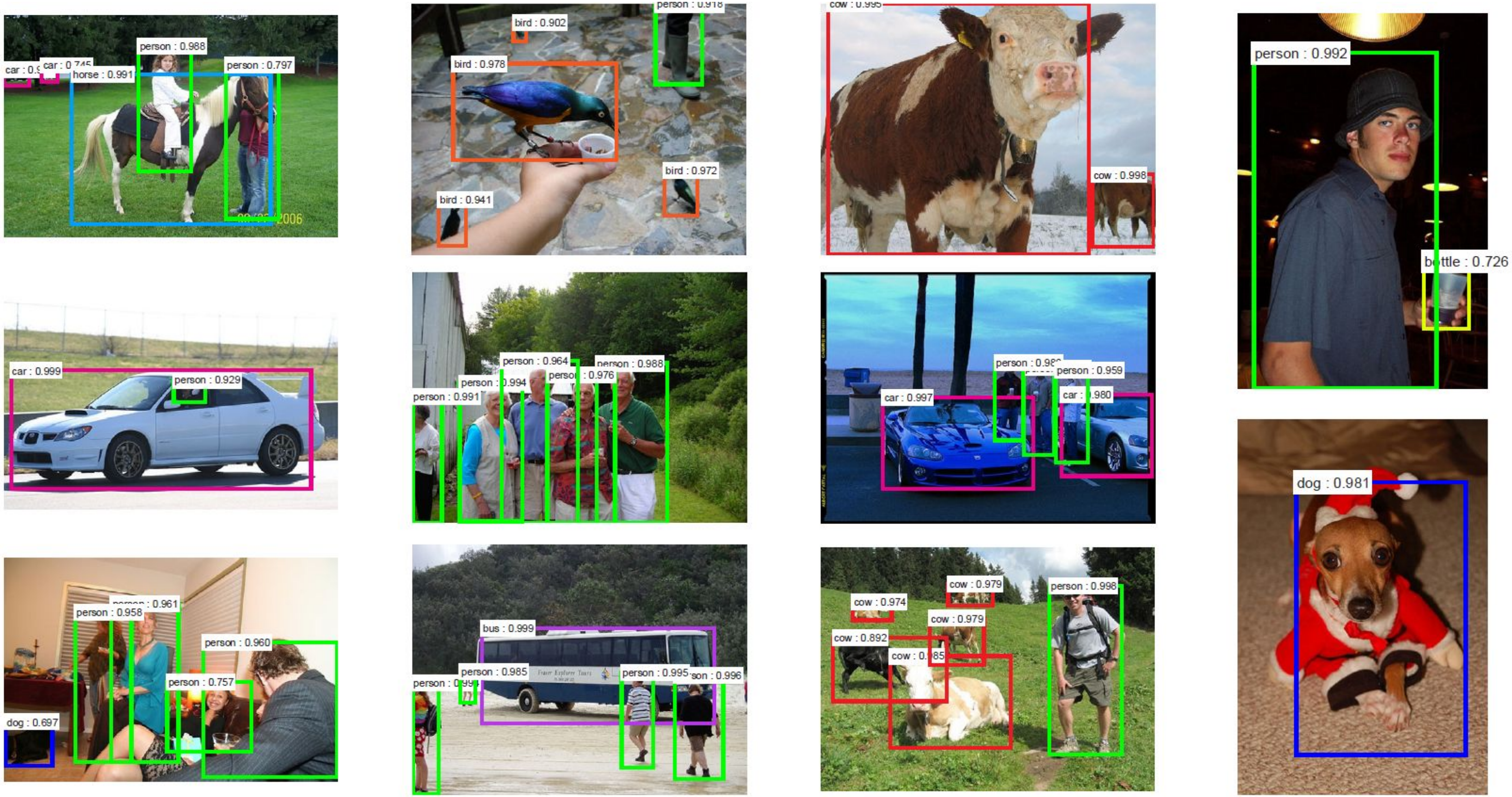
Experiments V: MS COCO (arXiv)

Table 11: Object detection results (%) on the **MS COCO** dataset. The model is VGG-16.

method	proposals	training data	COCO val		COCO test-dev	
			mAP@.5	mAP@[.5, .95]	mAP@.5	mAP@[.5, .95]
Fast R-CNN [2]	SS, 2000	COCO train	-	-	35.9	19.7
Fast R-CNN [impl. in this paper]	SS, 2000	COCO train	38.6	18.9	39.3	19.3
Faster R-CNN	RPN, 300	COCO train	41.5	21.2	42.1	21.5
Faster R-CNN	RPN, 300	COCO trainval	-	-	**42.7**	**21.9**

training data	2007 test	2012 test
VOC07	69.9	67.0
VOC07+12	73.2	-
VOC07++12	-	70.4
COCO (no VOC)	76.1	73.0
COCO+VOC07+12	**78.8**	-
COCO+VOC07++12	-	**75.9**

Qualitative Results



4. Summary

Summary

- Region Proposal Network sharing convolutional features with Object Detection Network makes region generation step nearly cost-free.

- Quality of proposals is improved with RPN wrt SS and EB.

- Object Detection system at 5-17 fps.

Summary

- Faster R-CNN is the basis of the winners of COCO and ILSVRC 2015 object detection competitions [1].

- RPN is also used in the winning entries of ILSVRC 2015 localization [1] and COCO 2015 segmentation competitions [2].

[1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” arXiv:1512.03385, 2015.
[2] J. Dai, K. He, and J. Sun, “Instance-aware semantic segmentation via multi-task network cascades,” arXiv:1512.04412, 2015.

Thank you !

Questions?