Object detection

Domonkos Varga

Budapest, 15th October, 2017

Questions

Please prepare to present an overview about object detection algorithms that use deep neural networks.

What are the key improvements from R-CNN to Fast R-CNN to Faster R-CNN?

Please compare Faster R-CNN to at least 2 more state-of-the-art detection algorithms and discuss the points scalability, compute efficiency and suitability for an automotive camera.

What method would you choose as a baseline for an automotive solution - and why?

Questions

Please prepare to present an overview about object detection algorithms that use deep neural networks.

What are the key improvements from R-CNN to Fast R-CNN to Faster R-CNN?

Please compare Faster R-CNN to at least 2 more state-of-the-art detection algorithms and discuss the points scalability, compute efficiency and suitability for an automotive camera.

What method would you choose as a baseline for an automotive solution - and why?

Overview and Brief History

Since 2012 when Alex Krizhevsky, Geoff Hinton, and Ilya Sutskever won ImageNet Convolutional Neural Networks became very popular and widely used for image classification tasks

= the result were surprising
top-1 and top-5 test set error rates of 37.5% and 17.0%
in comparison to results (47.1% and 28.2%)

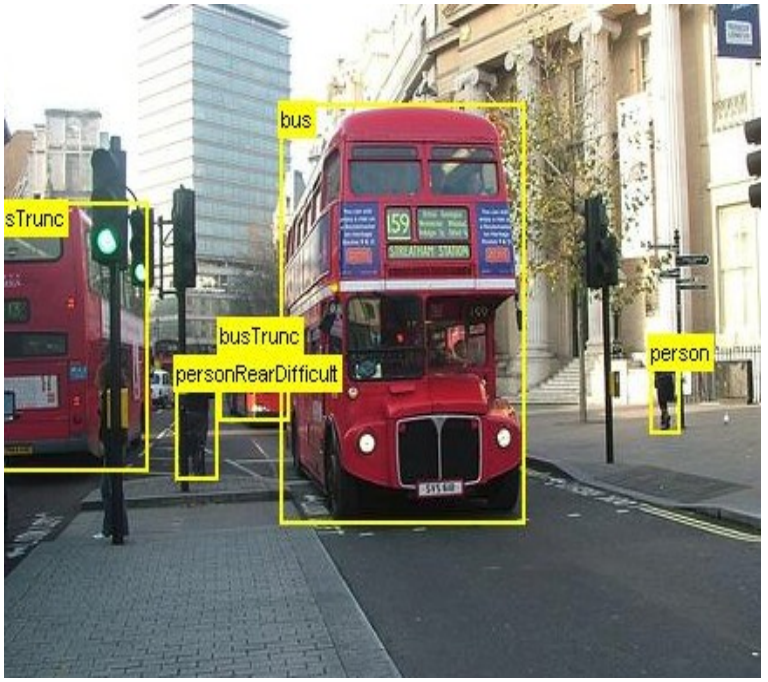
= it gave hope that CNN is the „right way“

= but the training time was enormous (days)

R-CNN

R-CNN = region proposal + CNN

- input image
- bounding box for the objects



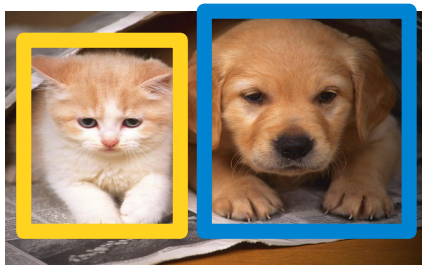
= Selective search (localization)

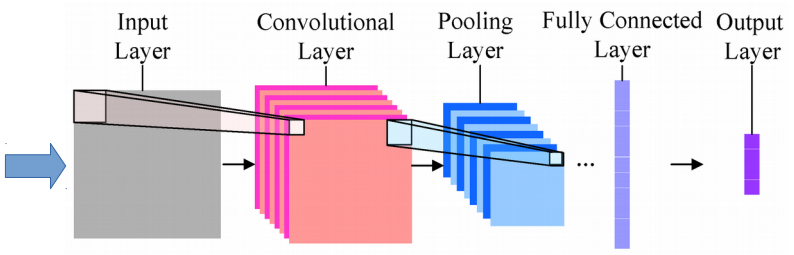
= Deep Learning CNN (feature extraction)

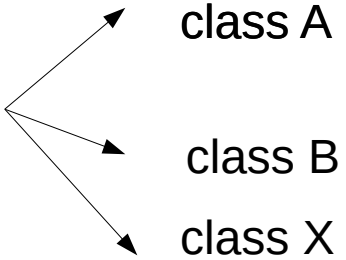
= Support Vector Machine (classification)

R-CNN: regions with CNN features

1) Input image







2) extract region proposals

3) CNN

4) classify regions

After the proposal R-CNN transforms the image to a standard size and passes it through a ConvNet.

Last layer can be a SVM which classifies the object type.

Final step is the tightening of the bounding box with a linear regressor.

R-CNN Training Steps

1) **Pretrain ConvNet** ← Imagenet

2) **Fine tune** it for object detection

3) **Cash feature vectors**

4) **Train Support Vector Machine** (ConvNet is fixed)

5) Train linear **bounding-box regressor** (Convnet is fixed)

Fast R-CNN

Fast R-CNN

- input: image + objects proposals
- output: bounding box for the objects

Improvements:

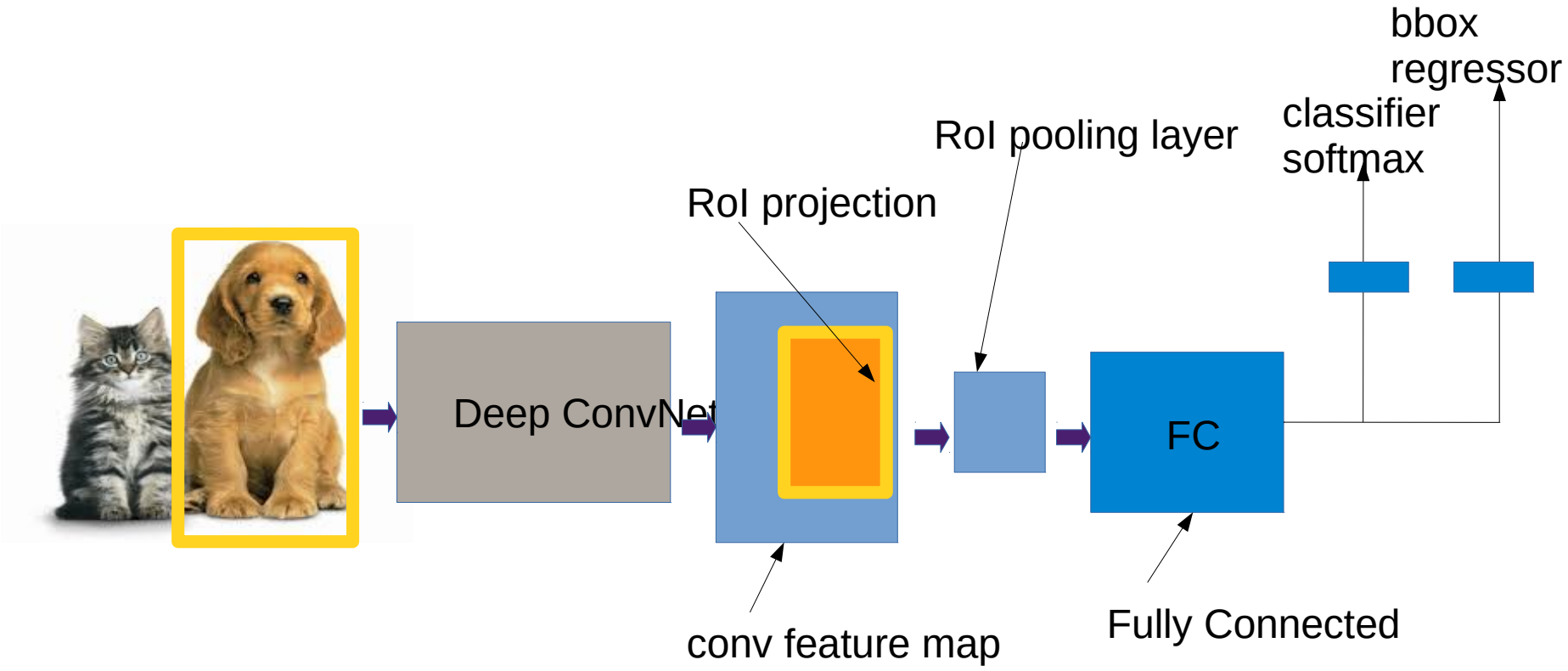
= higher detection quality (mAP)

= single step training with multitask loss function

= training updates all network layers

= no storage for feature cashing as in the case of R-CNN

Fast R-CNN architecture

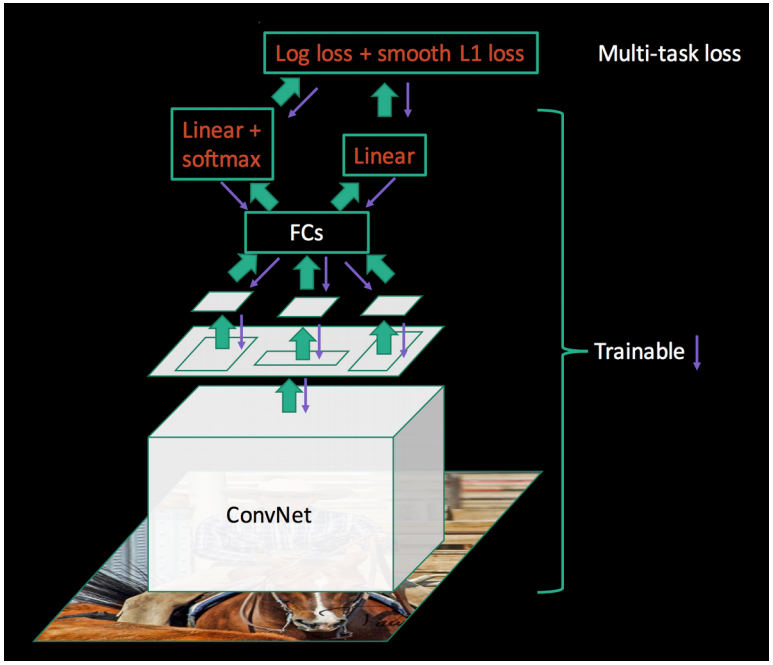


Excellent idea: Run the CNN just once per image and share the computation across proposals.

RoI (Region of Interest) layer. Only one pass/(image + proposals).

Feature extractor, classifier and bounding-box regressor are joined.

Faster R-CNN



Faster R-CNN

Problem: Selective search is rather slow

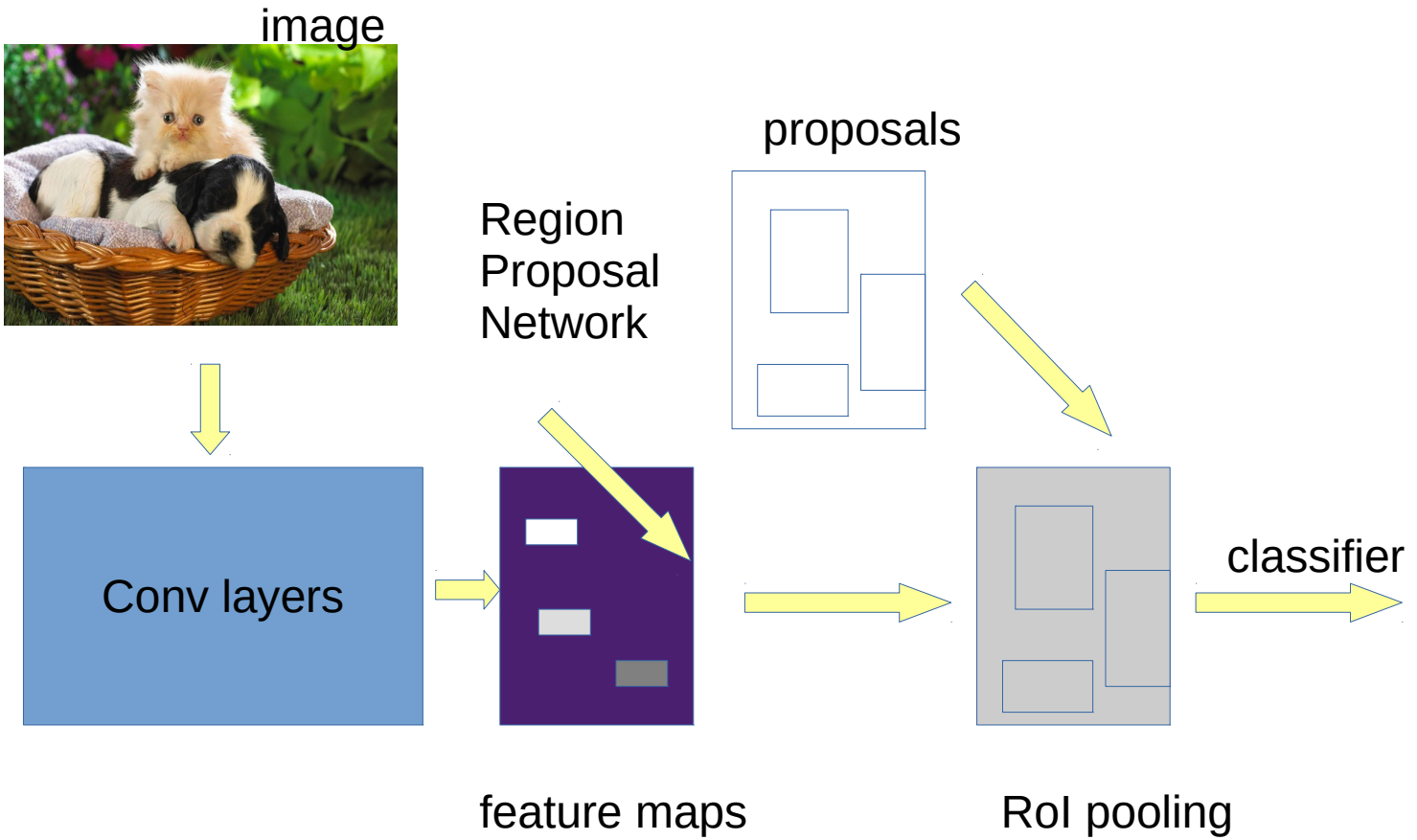
Improvement:

= Integration of computing prposals to ConvNet

= RPN (Region Proposal Network)

= RPN efficiently predicts region proposals with wide rage of scales and aspect rations (anchor boxes)

Faster R-CNN architecture

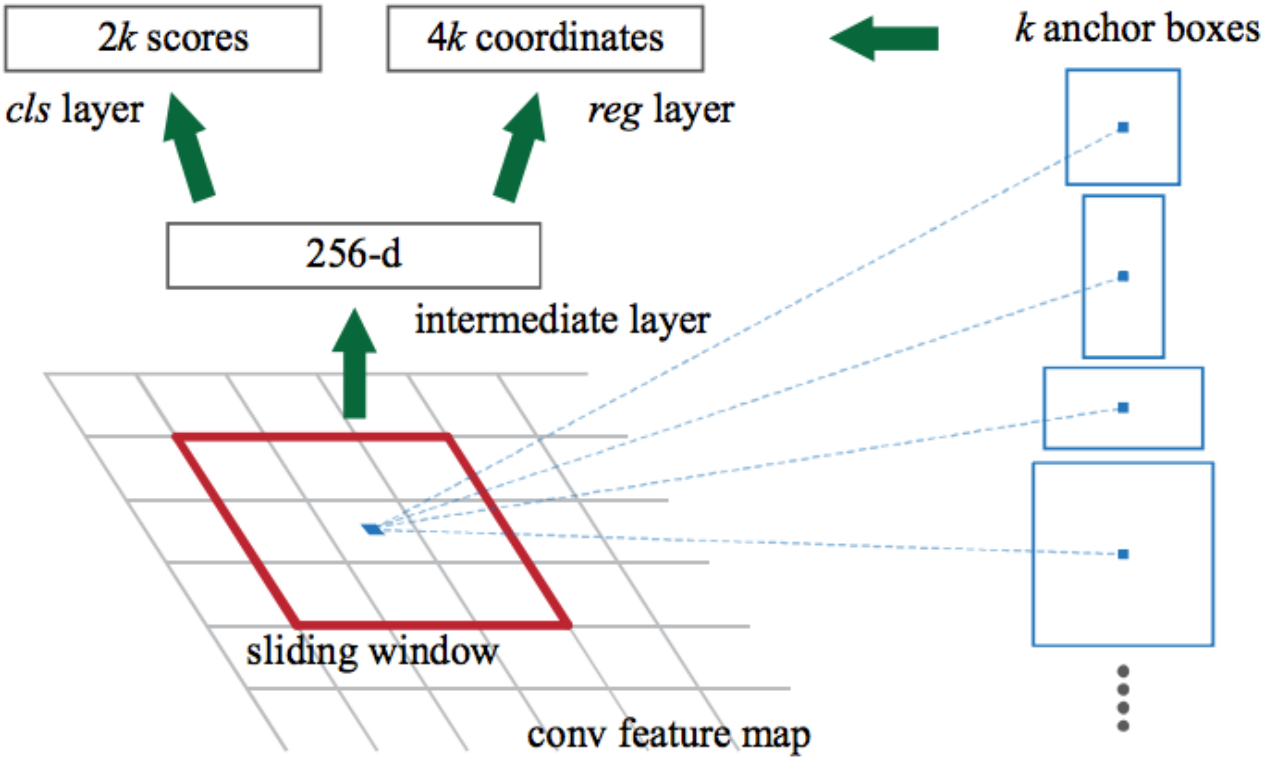


Idea: Integration of proposal system into the computing flow.

Sliding window for proposals.

Multiple region proposals at each sliding window location by changing scales and aspect ratios.

Faster R-CNN: anchor boxes



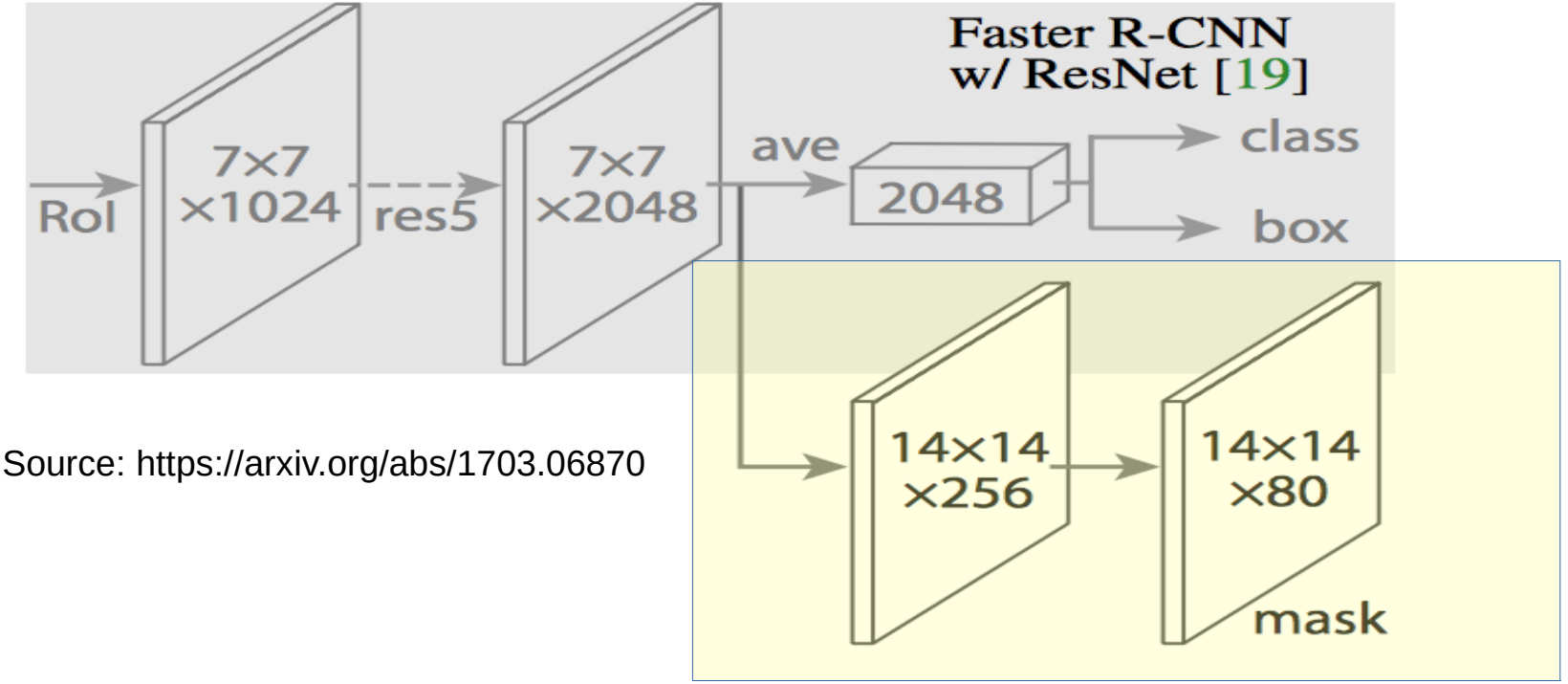
(source – S.Ren, K.He, R.Girschick, J.Sun: Faster R-CNN: Towards ...)

At each sliding window location *k* proposals.

For example 3 scales and 3 aspect ratios that is *k*=*9*.

Mask R-CNN architecture

It is worth mentioning although it is rather a segmentation.



Source: https://arxiv.org/abs/1703.06870

Idea: Extending Faster R-CNN object detection to pixelwise segmentation.

It adds a branch – a Fully Convolutional Network - to the Fast R-CNN flow which ouputs a binary mask whether a pixel belogs to the object or not.

Realigning of RoI pool is also necessary.

Questions

Please prepare to present an overview about object detection algorithms that use deep neural networks.

What are the key improvements from R-CNN to Fast R-CNN to Faster R-CNN?

Please compare Faster R-CNN to at least 2 more state-of-the-art detection algorithms and discuss the points scalability, compute efficiency and suitability for an automotive camera.

What method would you choose as a baseline for an automotive solution - and why?

R-CNN characteristics, problems

Although Region-based CNN achieves excellent object detection results it has drawbacks to overcome:

I) **Training is a multi-stage flow**.

a) ConvNet is trained on object proposal

b) SVM to ConvNet features

c) bounding-box regressor

II) **Training time is long**, days with GPU

III) **Object detection is slow**, 40-60 sec/image

Fast R-CNN improvements

Fast R-CNN has several contributions to R-CNN

I) On PASCAL VOC it has **higher detection quality** than R-CNN

II) In contrast to multi-stage R-CNN, Fast R-CNN **training is single-stage**

III) Training **updates all** network layers

IV) **No disk storage** is required (no feature cashing like R-CNN)

V) Fast R-CNN improves **training** *(9 x)* and **testing speed** *(213 x)*

Faster R-CNN improvements

Fast R-CNN is nearly real-time method if we do not take time spent on region proposal into account. Faster R-CNN improves this drawback.

I) Faster R-CNN **incorporates a Region Proposal Network**.

II) Region Proposal Network uses sliding-window with a so-called **anchor system** (scales and sizes) which ensures scalibility

III) It **increases testing speed**. 0.2 sec/image.

IV) It ensures a **better mAP** than Fast R-CNN on PASCAL VOC.

Comparison

	R-CNN	Fast R-CNN	Faster R-CNN
test time/image	50 sec	2 sec	0.2 ses
speedup	1 x	25 x	250 x
mAP (VOC)	~ 66%	~ 67%	~ 67%
learning	multistage	single stage	complex (4 steps)

Questions

Please prepare to present an overview about object detection algorithms that use deep neural networks.

What are the key improvements from R-CNN to Fast R-CNN to Faster R-CNN?

Please compare Faster R-CNN to at least 2 more state-of-the-art detection algorithms and discuss the points scalability, compute efficiency and suitability for an automotive camera.

What method would you choose as a baseline for an automotive solution - and why?

YOLO: You Only Look Once

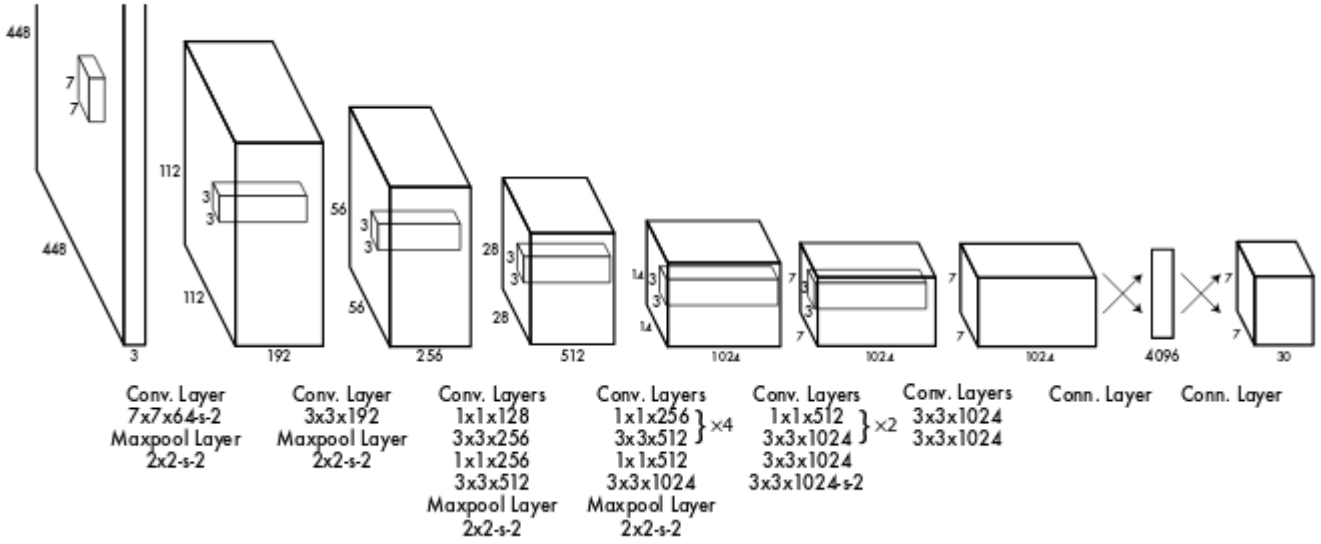
= Object detection is framed as regression problem to spatially separated bounding boxes and associated class probabilities.

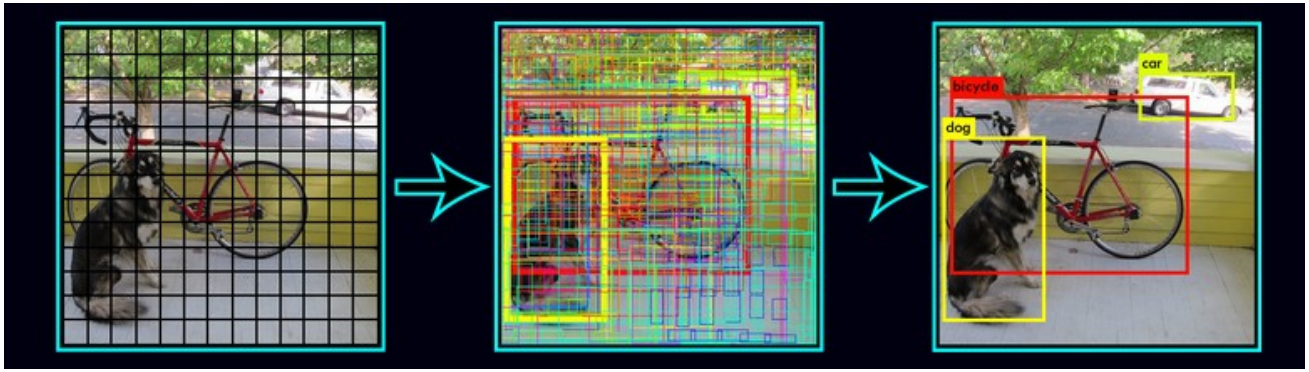
= First step is dividing the image into grid and predicting the probabilities of each bounding box.

= A single CNN predicts bounding boxes and class probabilities directly from full images in one evaluation.

= Like Faster R-CNN this method adjust prior bounding boxes.

YOLO: You Only Look Once





Source: https://pjreddie.com/darknet/yolo/

YOLO vs. Faster R-CNN

= YOLO is extremely fast, 10 times faster than Faster R-CNN.

= YOLO makes fewer background mistakes than Faster R-CNN.

= YOLO makes more localisation errors than Faster R-CNN.

= YOLO divides the image into a grid, while Faster R-CNN has RPN.

SPPNet vs. Faster R-CNN

= SPPNet introduces the Spatial Pyramid Pooling Layer which enables arbitrary sized images for the network.

= SPPNet is a flexible solution to handle different scales, sizes, and aspect ratios.

= SPPNet: 0.6 seconds per testing image, Faster R-CNN: 0.2 seconds per testing image.

= SPPNet: 60.9 % mAP (VOC), Faster R-CNN: 67 % mAP (VOC)

Scale invariance

two main ways of achieving scale invariant object detection:

1) „brute force“ approach: the network must learn scale-invariant object detection from the training data
(croppping; it may not contain the whole object,
warping; geometric distortions)

2) using image pyramid; after sampling an image at training pyramid scales are randomly chosen

[2] surprising result; single-scale detection performs almost as well as multi-scale detection (models trained and tested with either one or five scales)

Questions

Please prepare to present an overview about object detection algorithms that use deep neural networks.

What are the key improvements from R-CNN to Fast R-CNN to Faster R-CNN?

Please compare Faster R-CNN to at least 2 more state-of-the-art detection algorithms and discuss the points scalability, compute efficiency and suitability for an automotive camera.

What method would you choose as a baseline for an automotive solution - and why?

Industry

Although I also developed software for automative industry (I am also developing a project) I do not know deep enough the trends and requirements/visions of the big automotive firms.

1) When I had the task to implement a method in a short time, I would choose either YOLO or Faster R-CNN. They promise immediate results.

2) When I had the task to concentrate on a method, I would choose Fast R-CNN.
   a) This way I would separate the proposal and detector tasks.
   b) I had the chance to quickly adapt to new situation either changing, modifying, developing one of the modules.
   c) Maybe I could parallelize the proposal and detection tasks.
   d) Keeping track of „fashions“ is also important, not always the best and logical method wins a competition

References:

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik:
R-CNN for Object Detection
(UC Berkeley) Presentation, 2014

[2] Ross Girshick:
Fast R-CNN
arXiv:1504.08083

[3] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun:
Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks
arXiv:1506.01497

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick:
Mask R-CNN
arXiv:1703.06870

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun:
Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition
ArXiv:1406.4729

[6] P. Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, Yann LeCun:
OverFeat: Integrated Recognition, Localization and Detection using
Convolutional Networks
arXiv:1312.6229

References:

[7] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulder:
Selective search for object recognition
International Journal of Computer Vision (IJCV), 2013

[8] C. L. Zitnick and P. Doll´ar:
Edge boxes: Locating object proposals from edges
in European Conference on Computer Vision (ECCV), 2014

[9] Ross Girschick
Facebook AI Research
Training R-CNNs of various velocities
Presentation; Tools for Efficient Object Detection, ICCV 2015 Tutorial

[10] Fuxin Li
Fast, Faster and Mask R-CNN
Presentation; CS637

[11] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi
You Only Look Once: Unified, Real-Time Object Detection
arXiv:1506.02640