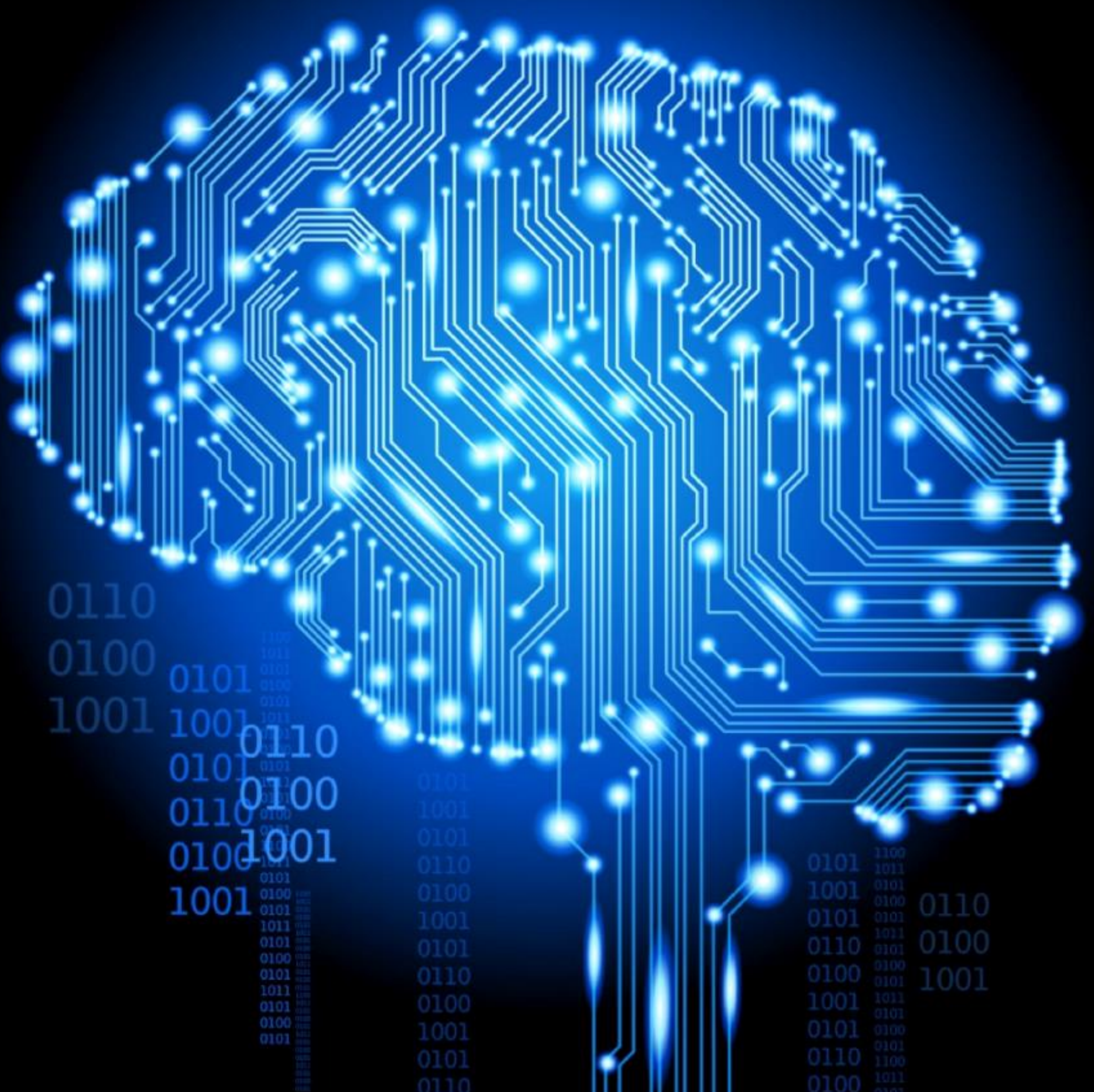
Deep Learning for object detection

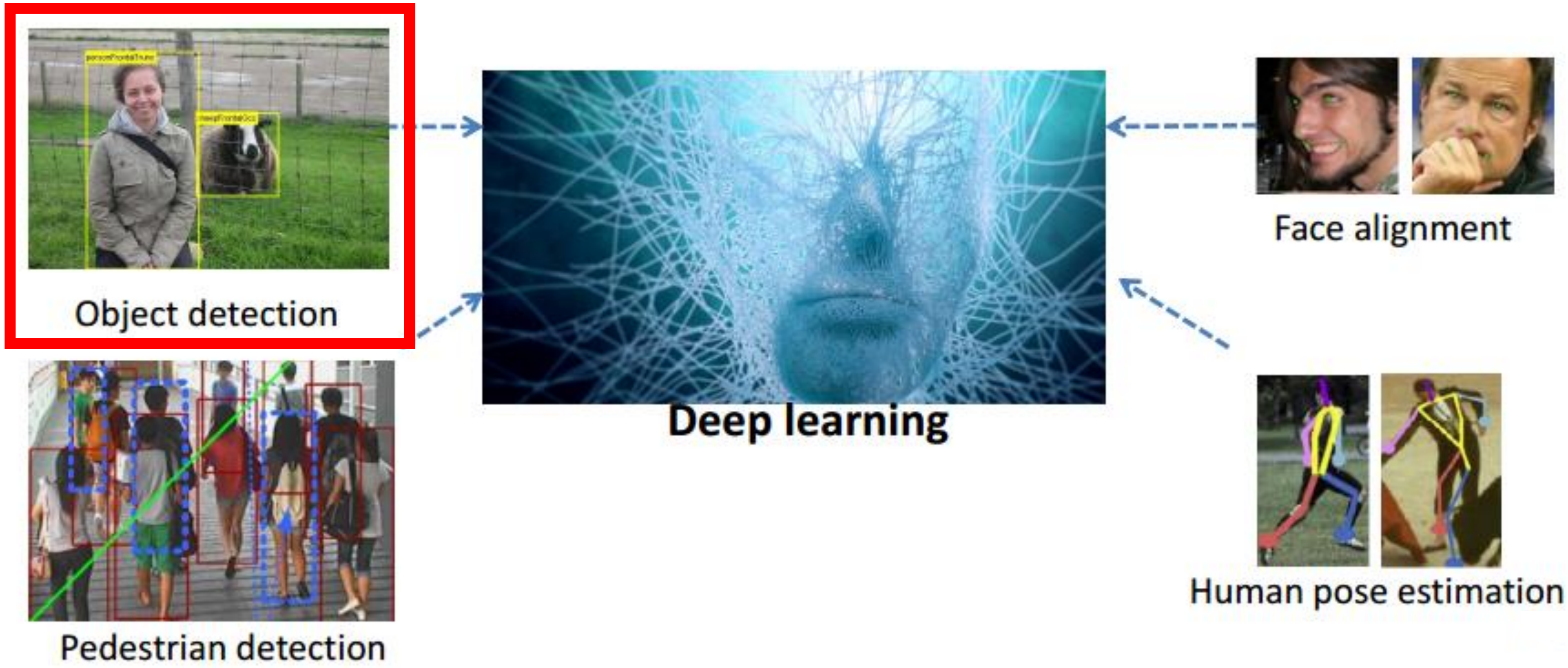
2016. 08. 01

Hyejeong Nam



Today topic

- Deep learning for object recognition
- Deep learning for object segmentation
- **Deep learning for object detection**
  - Pedestrian Detection
  - Human part localization
  - **General object detection**



http://www.isba2015.org/files/Deep_Learning_ISBA_2015.pdf

Today topic

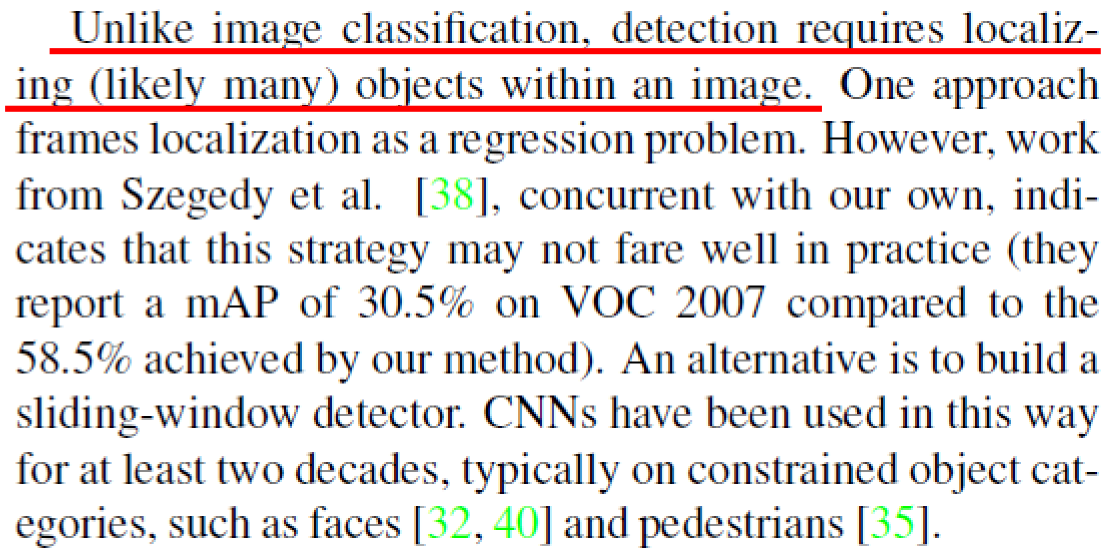
R-CNN → SPPnet → Fast-RCNN → Faster-RCNN
Framework Summary

R-CNN

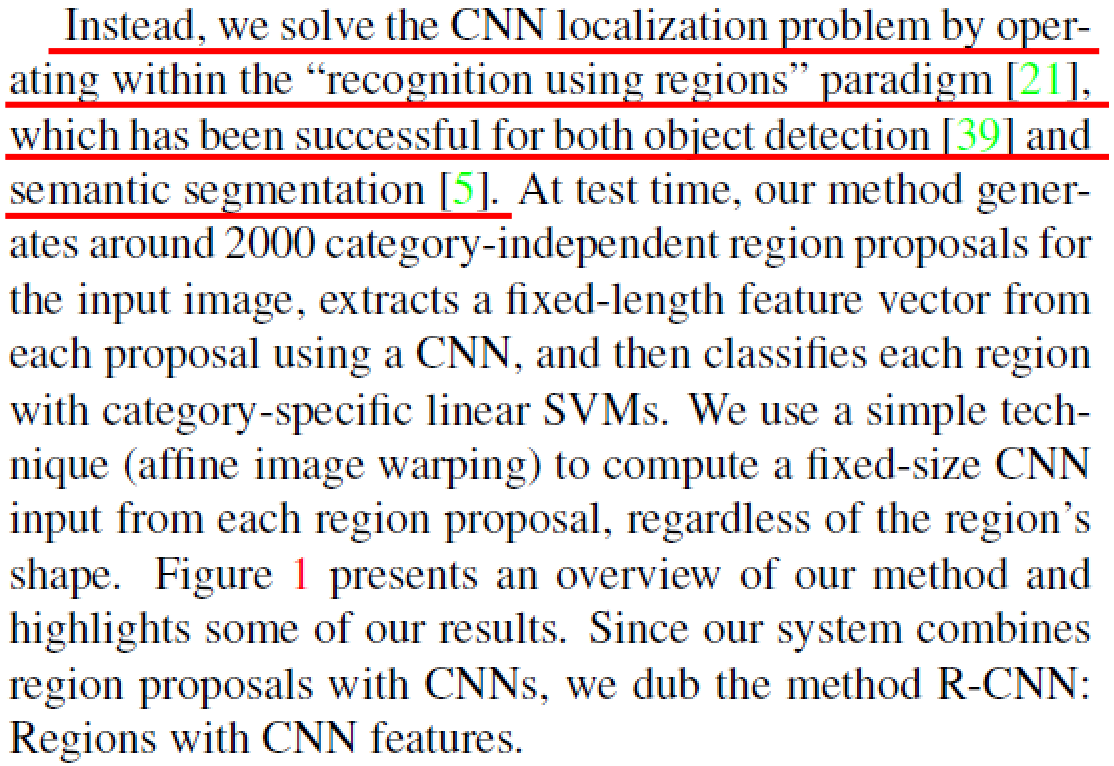
R-CNN
= Region-based Convolutional Neural Network

Detection

Classification



Unlike image classification, detection requires localizing (likely many) objects within an image. One approach frames localization as a regression problem. However, work from Szegedy et al. [38], concurrent with our own, indicates that this strategy may not fare well in practice (they report a mAP of 30.5% on VOC 2007 compared to the 58.5% achieved by our method). An alternative is to build a sliding-window detector. CNNs have been used in this way for at least two decades, typically on constrained object categories, such as faces [32, 40] and pedestrians [35].



Instead, we solve the CNN localization problem by operating within the "recognition using regions" paradigm [21], which has been successful for both object detection [39] and semantic segmentation [5]. At test time, our method generates around 2000 category-independent region proposals for the input image, extracts a fixed-length feature vector from each proposal using a CNN, and then classifies each region with category-specific linear SVMs. We use a simple technique (affine image warping) to compute a fixed-size CNN input from each region proposal, regardless of the region's shape. Figure 1 presents an overview of our method and highlights some of our results. Since our system combines region proposals with CNNs, we dub the method R-CNN: Regions with CNN features.

R-CNN

Rich feature hierarchies for accurate object detection and semantic segmentation
Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik / **CVPR 2014**

In the paper, a simple and scalable detection algorithm is proposed.
Two key points:

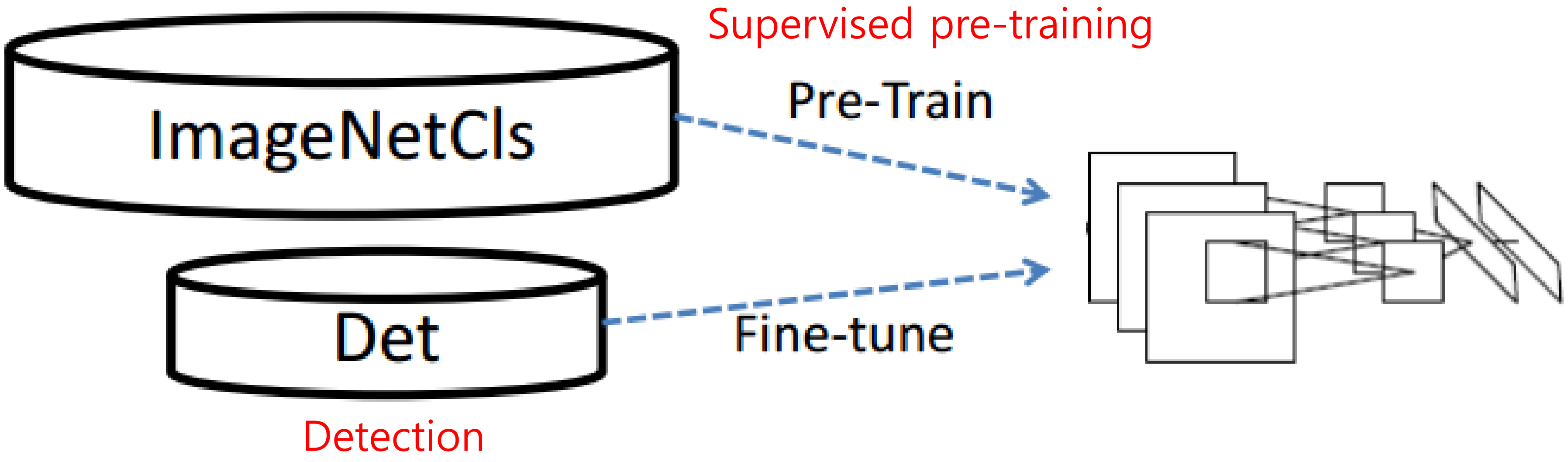
1. **localizing objects with a deep network**
2. **training a high-capacity model
with a small quality of annotated detection data**

R-CNN

Rich feature hierarchies for accurate object detection and semantic segmentation

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik / **CVPR 2014**

- Pretrain for the 1000-way ILSVRC image classification task (1.2 million images)
- Fine-tune the CNN for detection
  - Transfer the representation learned from ILSVRC Classification to PASCAL (or ImageNet) detection



Supervised pre-training

ImageNetCls

Pre-Train

Det

Fine-tune

Detection

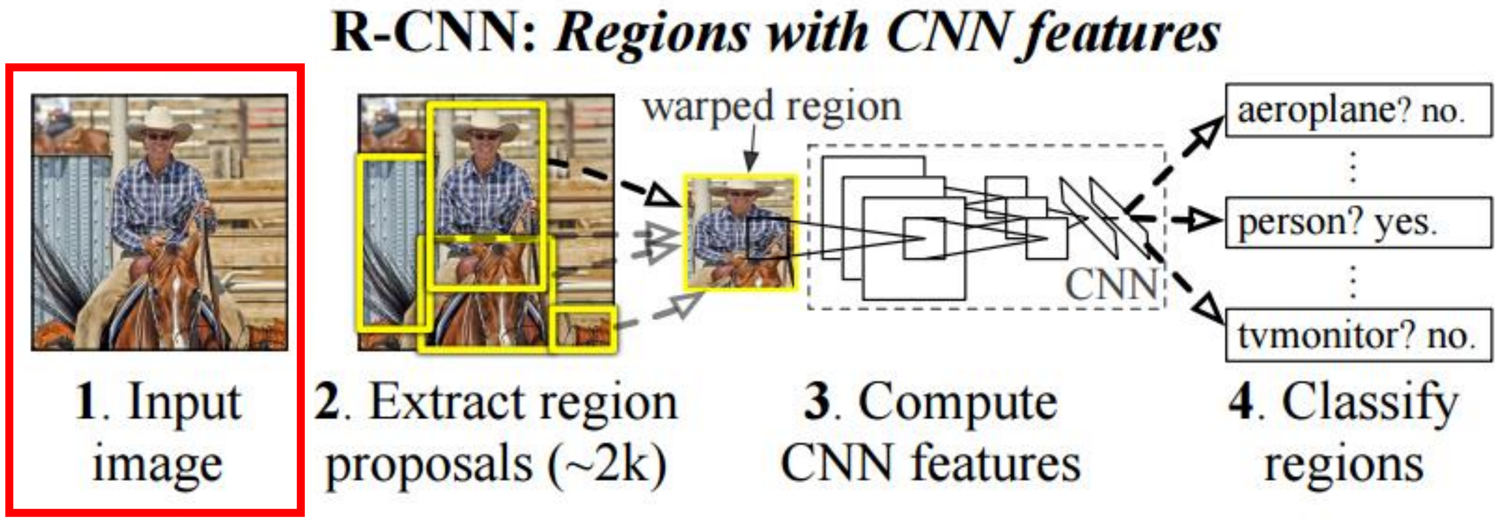
Network from Krizhevsky, Sutskever & Hinton. NIPS 2012 Also called “AlexNet”

R-CNN

Rich feature hierarchies for accurate object detection and semantic segmentation

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik / **CVPR 2014**

Object detection system overview



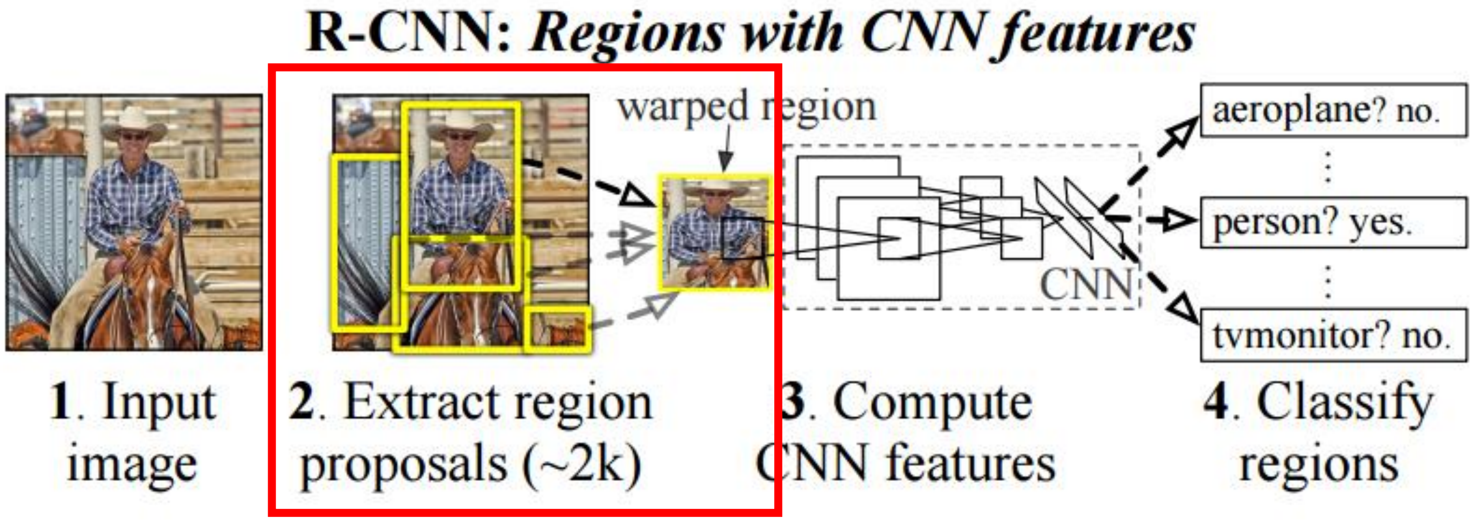
1. Input image : Nothing special to say, any images as inputs.

R-CNN

Rich feature hierarchies for accurate object detection and semantic segmentation

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik / **CVPR 2014**

Object detection system overview



2. Extract region proposals : extracts (around 2000 bottom-up) region proposals & generates category-independent region proposals

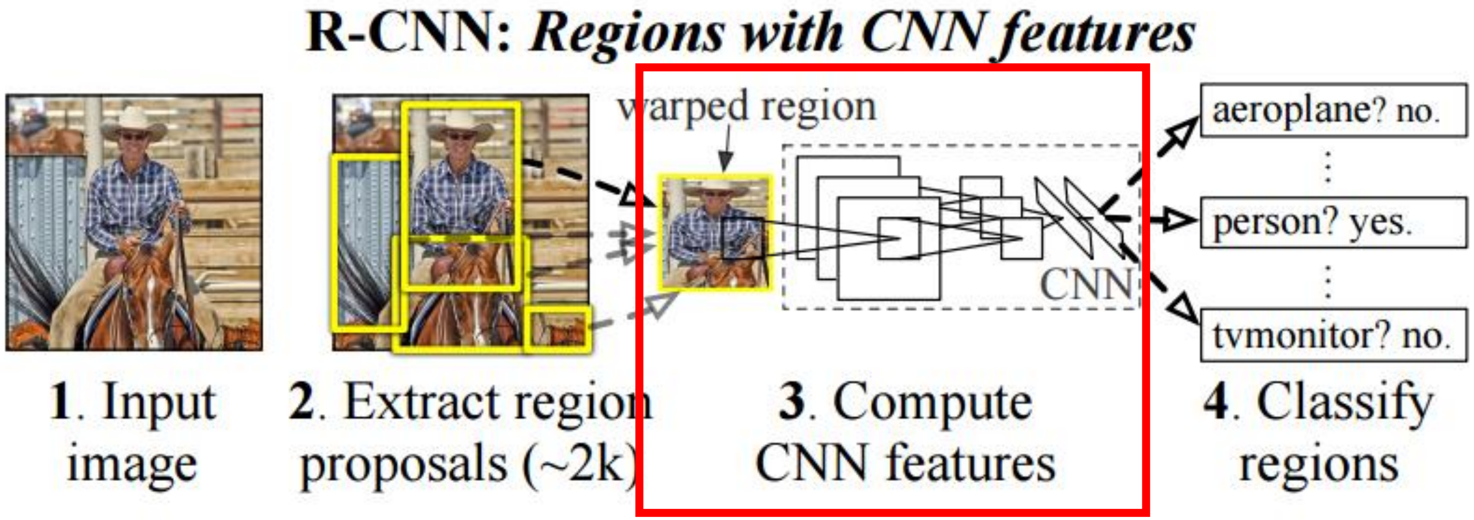
(using selective search's "fast mode" in all experiments)

R-CNN

Rich feature hierarchies for accurate object detection and semantic segmentation

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik / UC Berkeley / CVPR 2014

Object detection system overview



3. Compute CNN features : 5 Convolution Layers, 2 Fully Connected layers
Any architecture, not just AlexNet

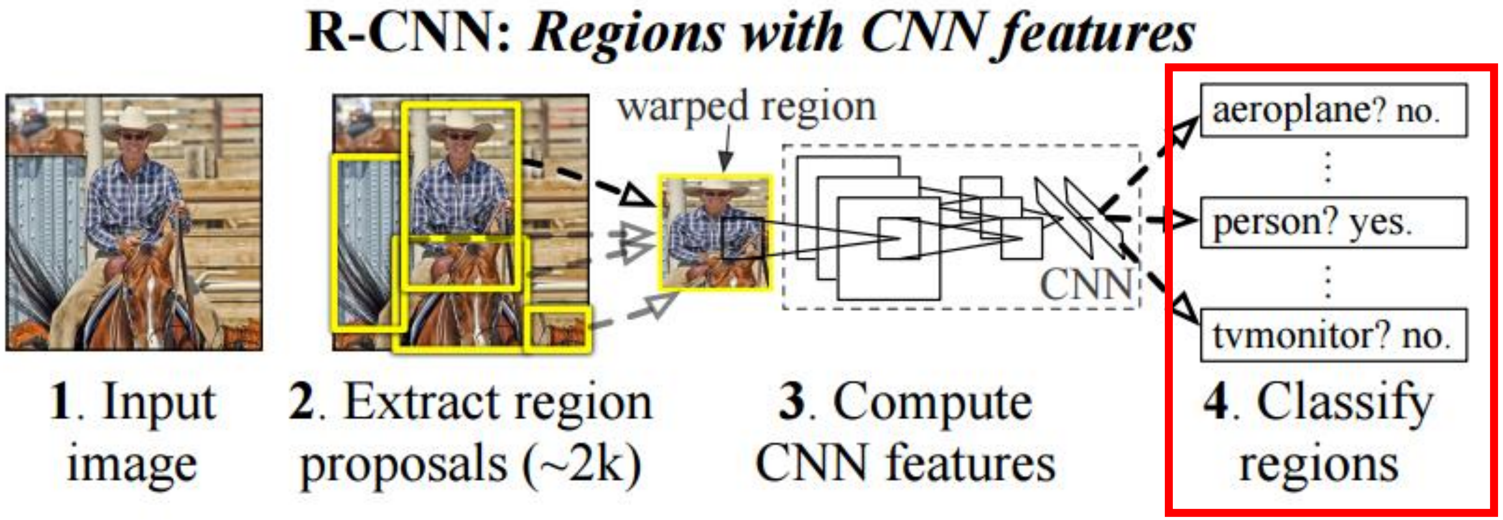
Extracts a fixed-length feature vector from each region

R-CNN

Rich feature hierarchies for accurate object detection and semantic segmentation

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik / **CVPR 2014**

Object detection system overview



4. Classify Regions : score each extracted feature vector using the linear SVM trained for that class.

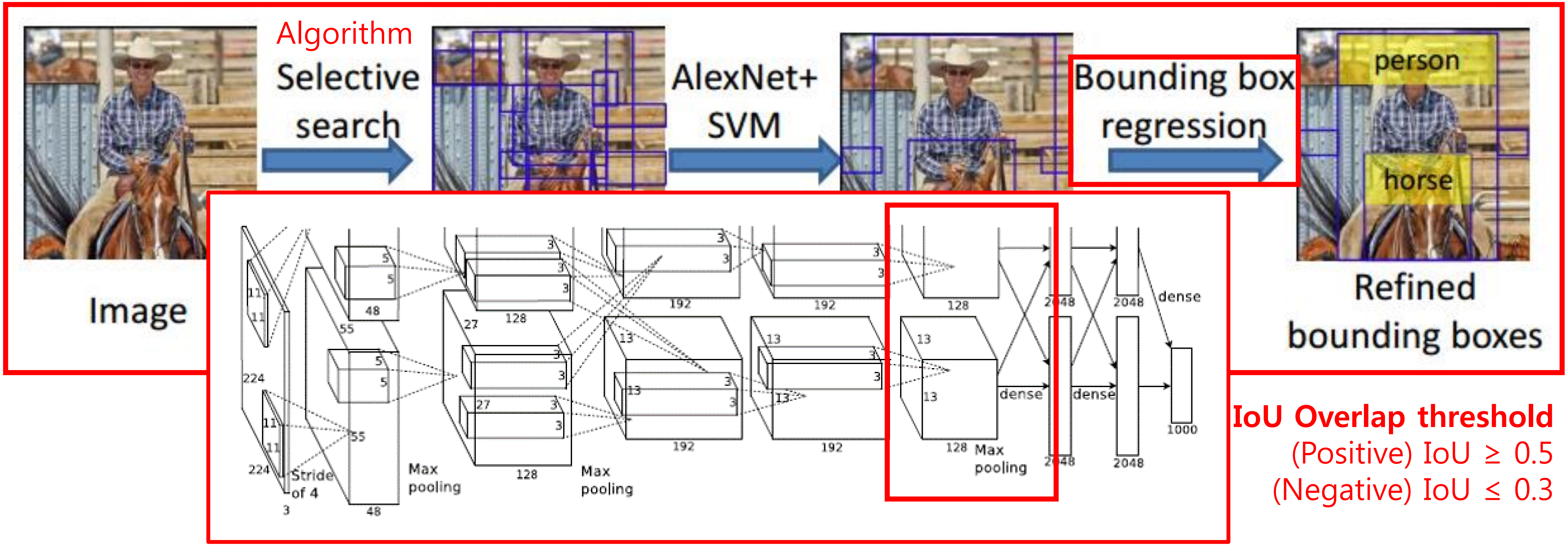
R-CNN

Rich feature hierarchies for accurate object detection and semantic segmentation

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik / **CVPR 2014**

Object detection system overview

Detection system



Algorithm

Selective search

AlexNet+ SVM

Bounding box regression

Image

Refined bounding boxes

IoU Overlap threshold

(Positive) IoU ≥ 0.5

(Negative) IoU ≤ 0.3

R-CNN

Rich feature hierarchies for accurate object detection and semantic segmentation

Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik / **CVPR 2014**

Why efficient?

1. CNN shared features (parameters), original characteristic of CNN
2. CNN low-dimensional computing. (Reduced number of features extracted)

Conclusion

Achieved performance through:

1. apply high-capacity CNN works to bottom-up region proposals in order to localize and segment objects.
2. paradigm for training large CNNs when labeled training data is scarce

R-CNN

R-CNN -> SPPnet -> Fast-RCNN -> Faster-RCNN

- Advantage : Good performance level
- Weakness :
  - When region proposal was classified in CNN, performance level is low due to image modify/loss.
  - Velocity is low because CNN computation was practiced after classifying region proposals.
  - Algorithm using in region proposal is not beneficial for fast-calculation on GPU.

SPPnet solve

Fast-RCNN solve

Faster-RCNN solve

SPPnet

SPPnet
= Spatial Pyramid Pooling in
Deep Convolutional Networks for Visual Recognition

Why?

- Existing deep convolutional neural networks (CNNs) require a fixed-size (e.g., 224×224) input image. This requirement is "artificial" and may reduce the recognition accuracy for the images or sub-images of an arbitrary size/scale.

- To eliminate the above requirement, research team equips the networks with another pooling strategy, "spatial pyramid pooling"

SPPnet

Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun / **ECCV2014**

CNN → convolutional layers + fully-connected layers

do not require a fixed image size and
can generate feature maps of any sizes

need to have fixed size/length input
by their definition

∴ **the fixed size constraint comes only from the fully-connected layers,
which exist at a deeper stage of the network**

Key Point

→ introduce a spatial pyramid pooling (SPP) layer to remove the fixed-size constraint of the network

SPPnet

Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun / **ECCV2014**

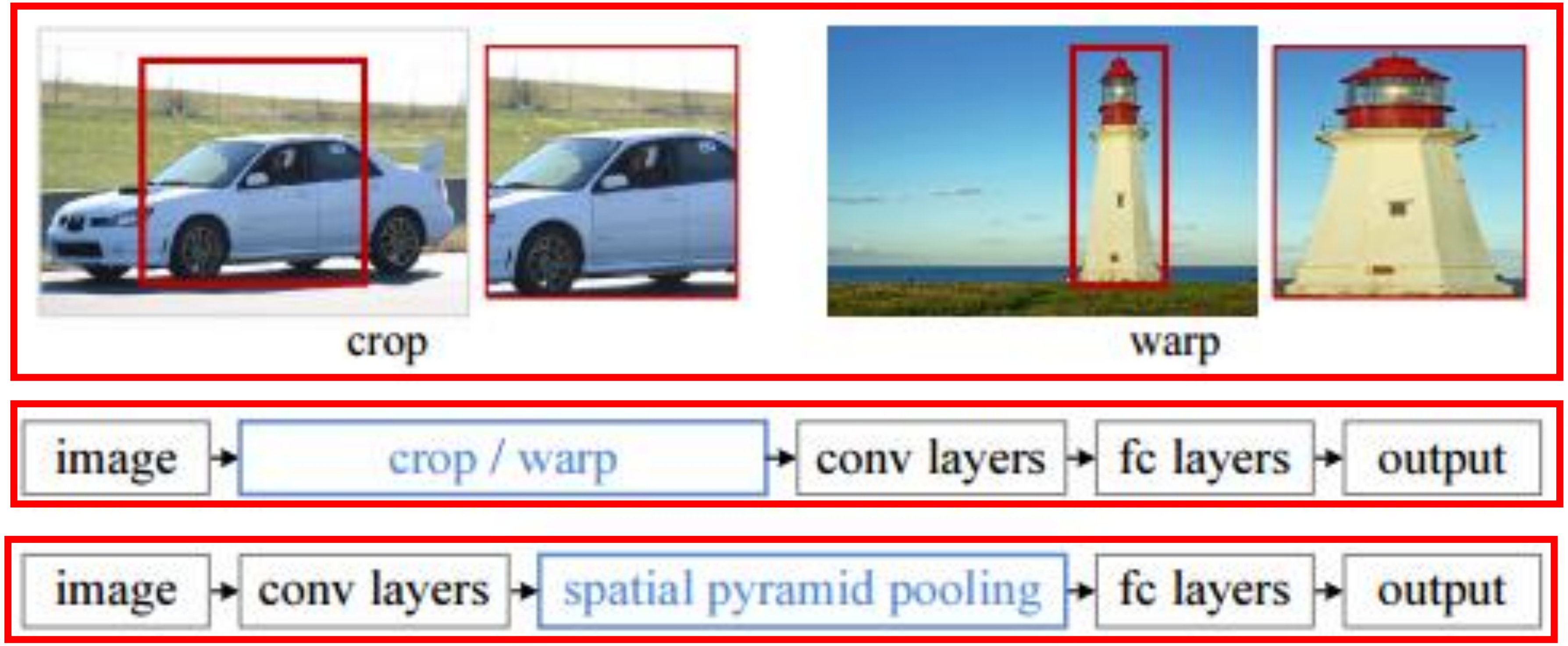


image → crop / warp → conv layers → fc layers → output

image → conv layers → spatial pyramid pooling → fc layers → output

Cropping or warping to fit a fixed size

A conventional CNN

SPPnet structure

↑ SPP layer on top of the last convolutional layer

SPPnet

Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

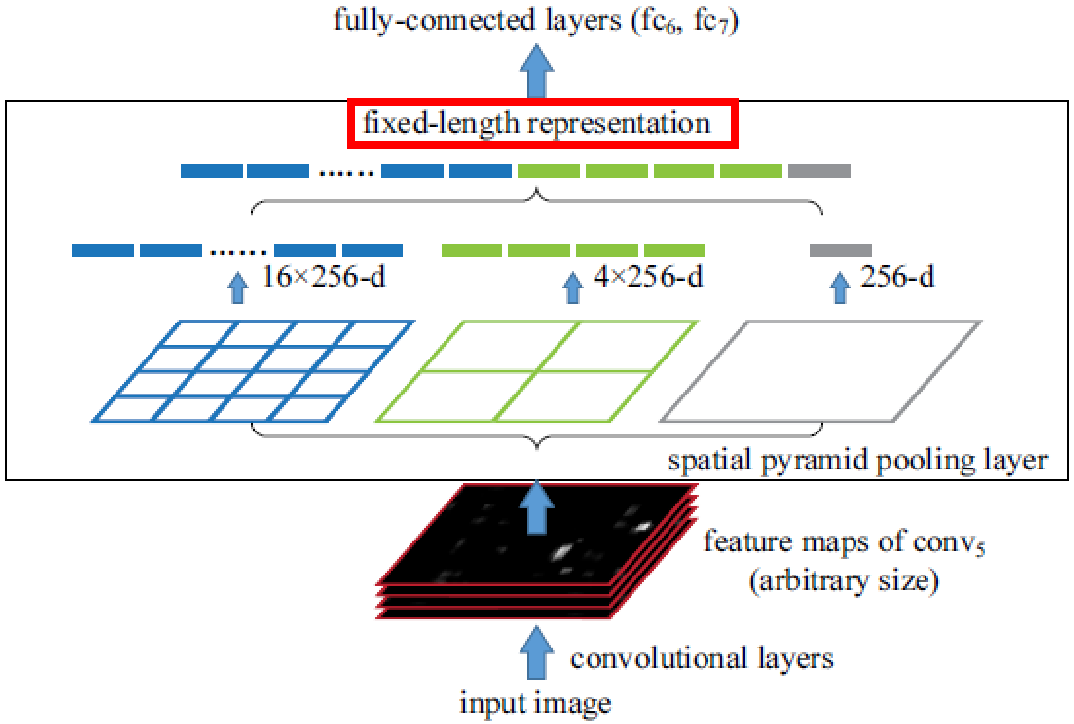
Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun / **ECCV2014**

SPPnet

- a new network structure

- Classification : improves all CNNs
- Detection : 20-60x faster than R-CNN, as accurate

- variable input size/scale
  - multi-size training
  - multi-scale testing
  - full-image view
- multi-level pooling
  - robust to deformation
- operates on feature maps
  - pooling in regions



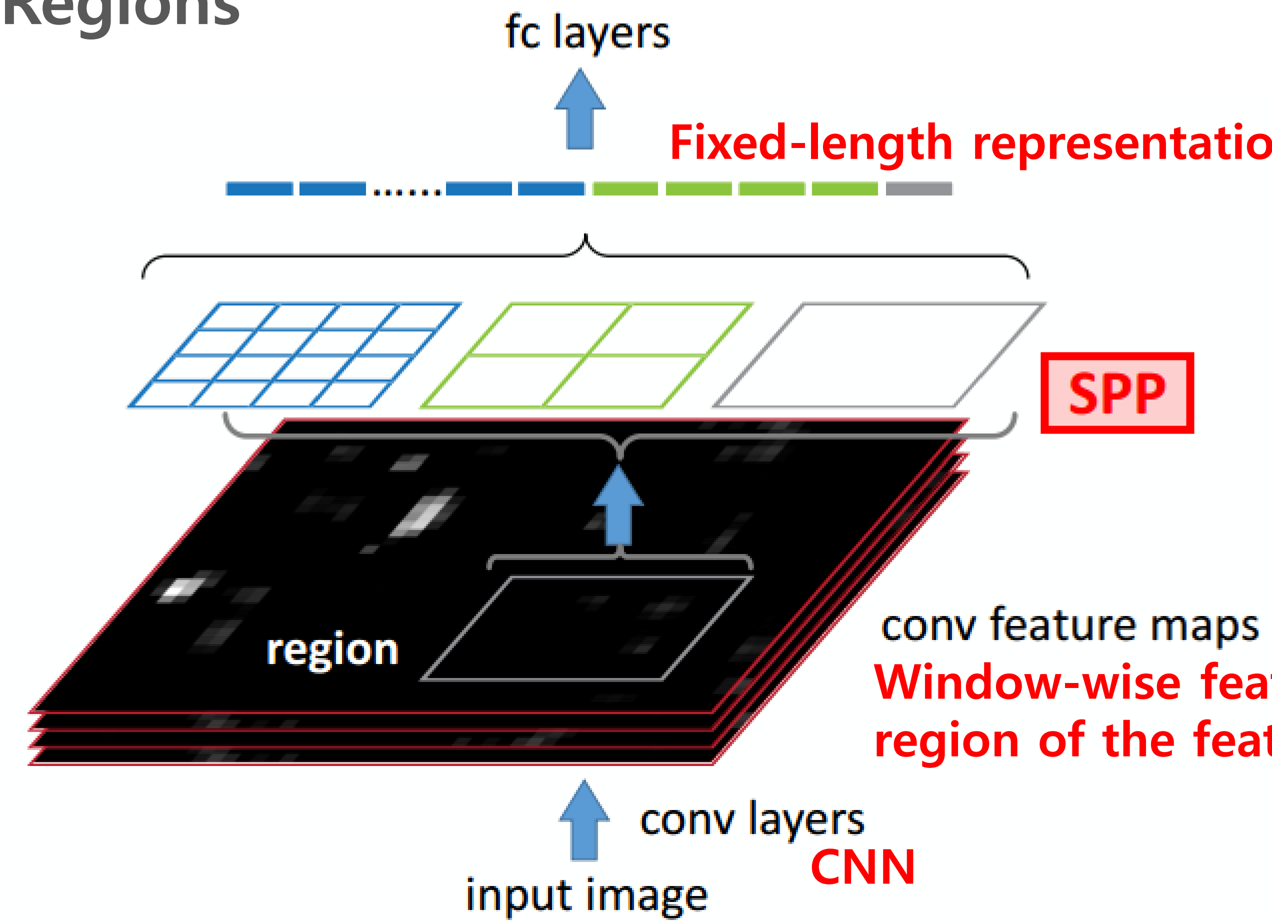
http://kaiminghe.com/eccv14sppnet/sppnet_ilsvrc2014.pdf

SPPnet

Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun / **ECCV2014**

Detection: SPP on Regions



Fixed-length representation

SPP

Window-wise features from region of the feature maps

CNN

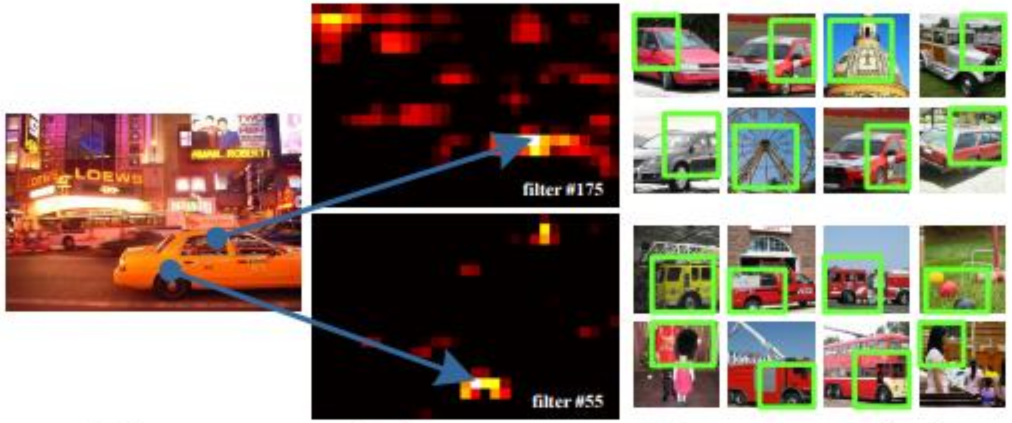
http://kaiminghe.com/eccv14sppnet/sppnet_ilsvrc2014.pdf

SPPnet

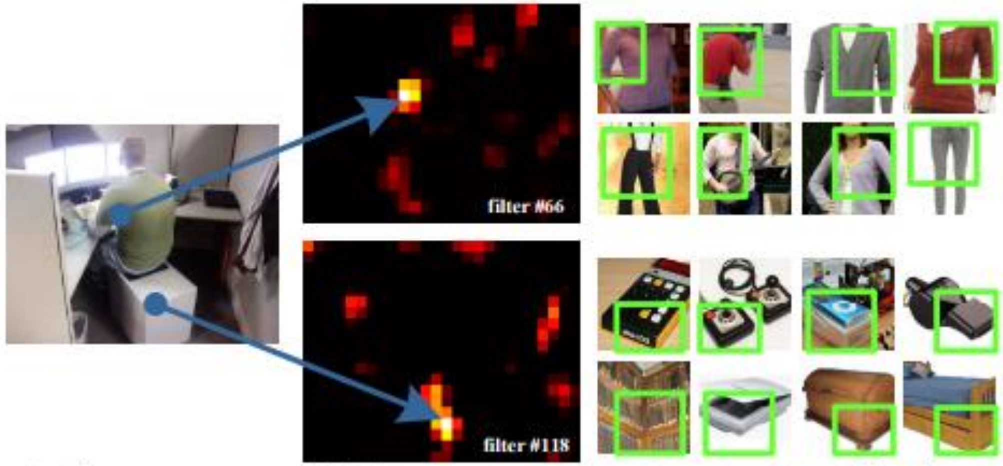
Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition (2014)
Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun / **ECCV2014**

Visualization of the feature maps

(a) Two images in Pascal VOC 2007.
(b) The feature maps of some conv5 filters. The arrows indicate the strongest responses and their corresponding positions in the images.
(c) The ImageNet images that have the strongest responses of the corresponding filters. The green rectangles mark the receptive fields of the strongest responses.



(a) image (b) feature maps (c) strongest activations



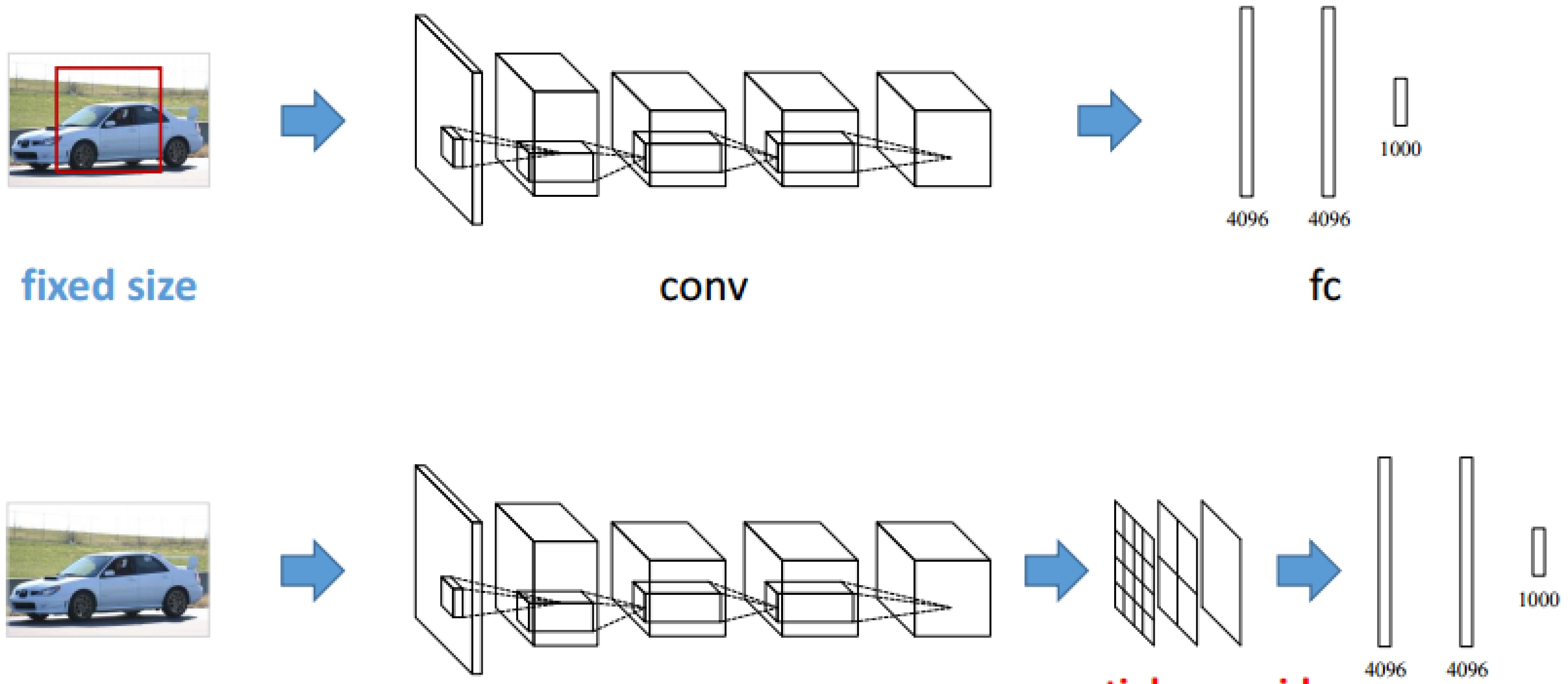
(a) image (b) feature maps (c) strongest activations

http://kaiminghe.com/eccv14sppnet/sppnet_ilsvrc2014.pdf

SPPnet

Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun / **ECCV2014**



traditional CNN

fixed size

conv

fc

SPP-net

any size

spatial pyramid pooling

- Fix bin numbers
- **DO NOT** fix bin size

http://kaiminghe.com/eccv14sppnet/sppnet_ilsvrc2014.pdf

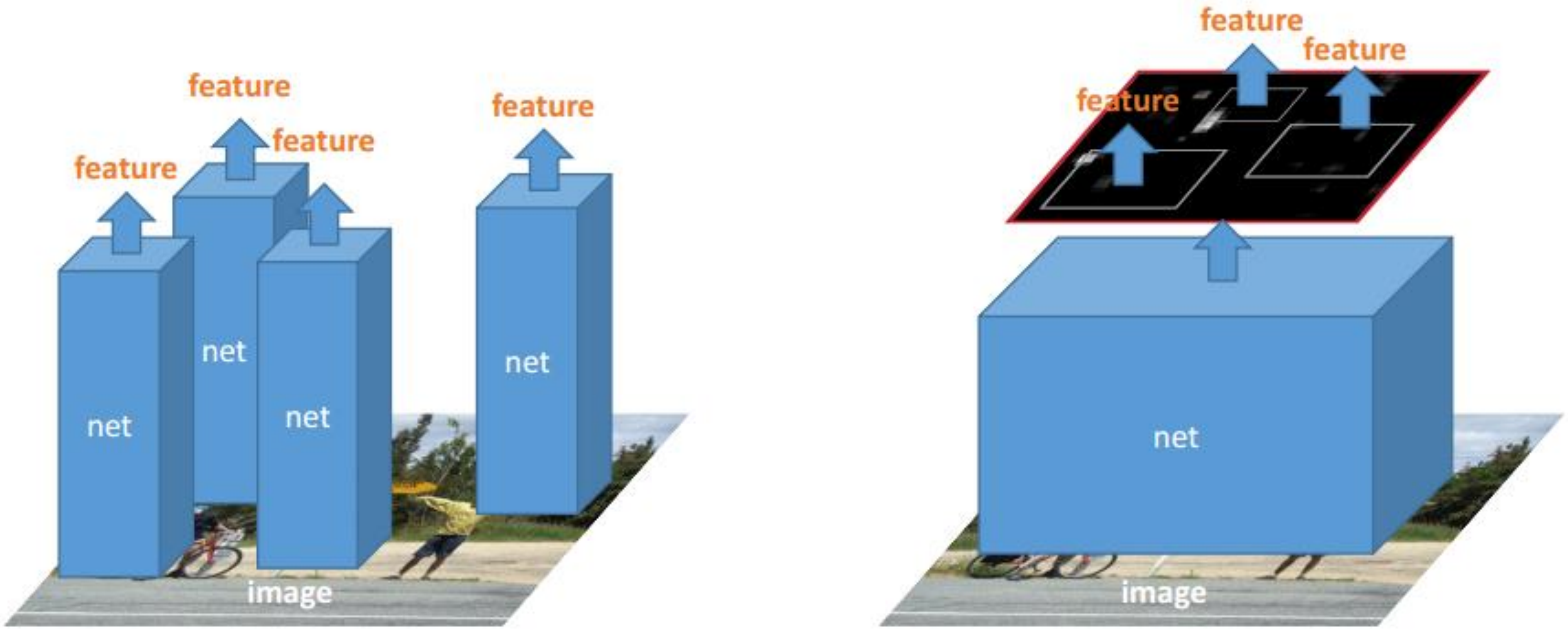
SPPnet

Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition (2014)

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun / **ECCV2014**

R-CNN vs. SPPnet

image regions vs. feature map regions



R-CNN
2000 nets on image regions

SPP-net
1 net on full image

http://kaiminghe.com/eccv14sppnet/sppnet_ilsvrc2014.pdf

SPPnet

R-CNN -> SPPnet -> Fast-RCNN -> Faster-RCNN

- ~~When region proposal was classified in CNN, performance level is low due to image modify/loss.~~
- ~~Velocity is low because CNN computation was practiced after classifying region proposals.~~
- Algorithm using in region proposal is not beneficial for fast-calculation on GPU.

Fast-RCNN solve

Faster-RCNN solve

Fast R-CNN

Fast R-CNN
= Fast Region-based Convolutional Networks
for object detection

Fast R-CNN is a fast framework for object detection with deep ConvNets. Fast R-CNN

- trains state-of-the-art models, like VGG16, **9x faster than traditional R-CNN and 3x faster than SPPnet,**
- **runs 200x faster than R-CNN and 10x faster than SPPnet at test-time,**
- has a significantly higher mAP on PASCAL VOC than both R-CNN and SPPnet,
- and is written in Python and C++/Caffe

https://github.com/rbgirshick/fast-rcnn

Fast R-CNN

Fast R-CNN

Ross Girshick / **ICCV2015**

R-CNN, however, has notable drawbacks:

1. **Training is a multi-stage pipeline.**

   R-CNN first finetunes a ConvNet on object proposals using log loss. Then, it fits SVMs to ConvNet features. These SVMs act as object detectors, replacing the softmax classi- fier learnt by fine-tuning. In the third training stage, bounding-box regressors are learned.

2. **Training is expensive in space and time.**

   For SVM and bounding-box regressor training, features are extracted from each object proposal in each image and written to disk. With very deep networks, such as VGG16, this process takes 2.5 GPU-days for the 5k images of the VOC07 trainval set. These features require hundreds of gigabytes of storage.

3. **Object detection is slow.**

   At test-time, features are extracted from each object proposal in each test image. Detection with VGG16 takes 47s / image (on a GPU).

Fast R-CNN

Fast R-CNN
Ross Girshick / **ICCV2015**

Four key points:

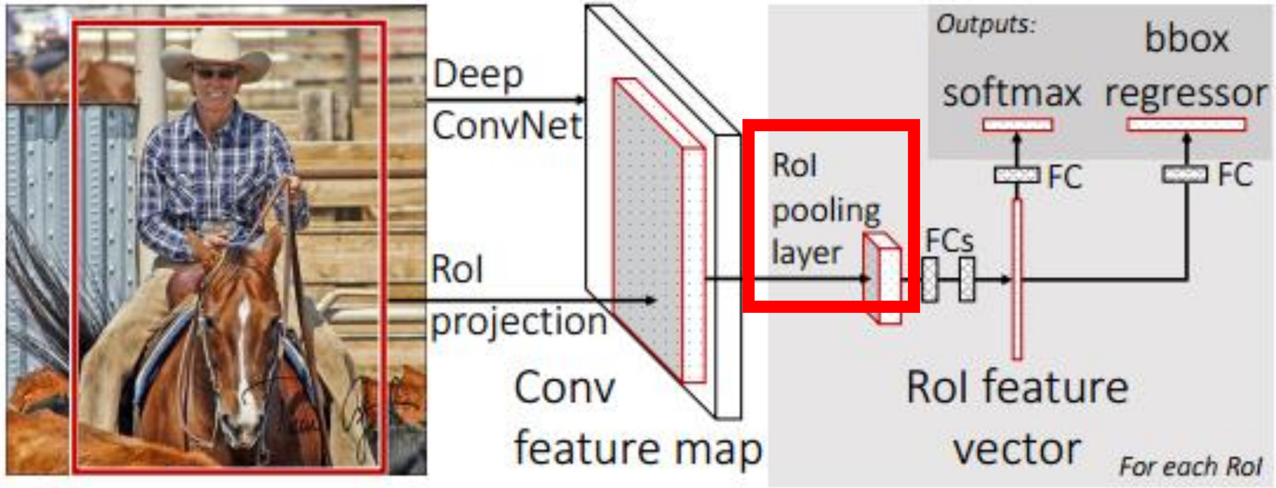
1. **Higher detection quality (mAP) than R-CNN, SPPnet**
2. **Training is single-stage, using a multi-task loss**
3. **Training can update all network layers**
4. **No disk storage is required for feature caching**

Fast R-CNN

Fast R-CNN

Ross Girshick / **ICCV2015**

Fast R-CNN architecture and training



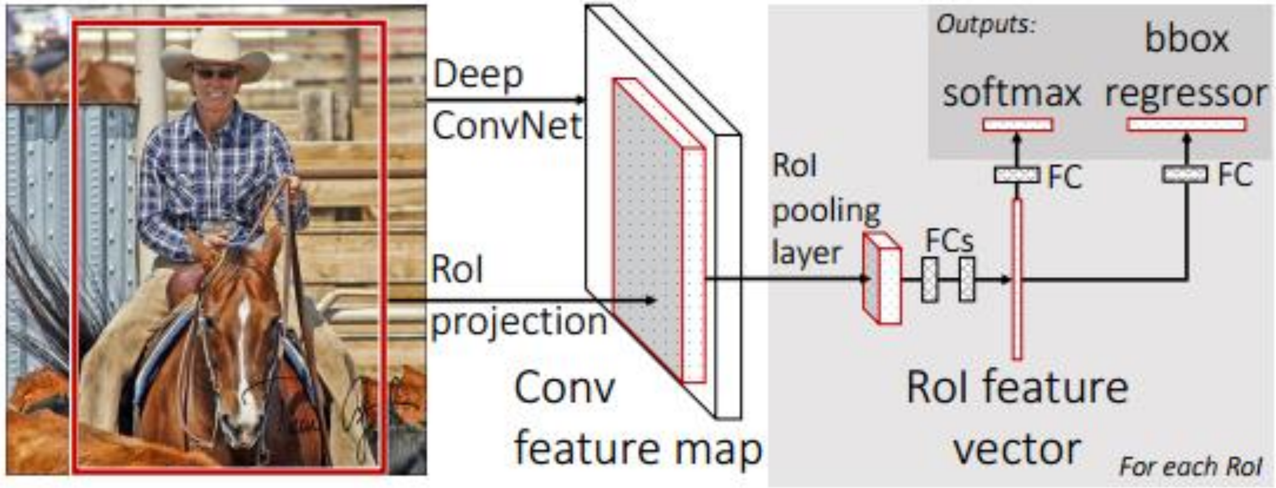
* RoI pooling layer = for fixing output vector size when input image size is various

Fast R-CNN

Fast R-CNN

Ross Girshick / **ICCV2015**

Fast R-CNN architecture and training

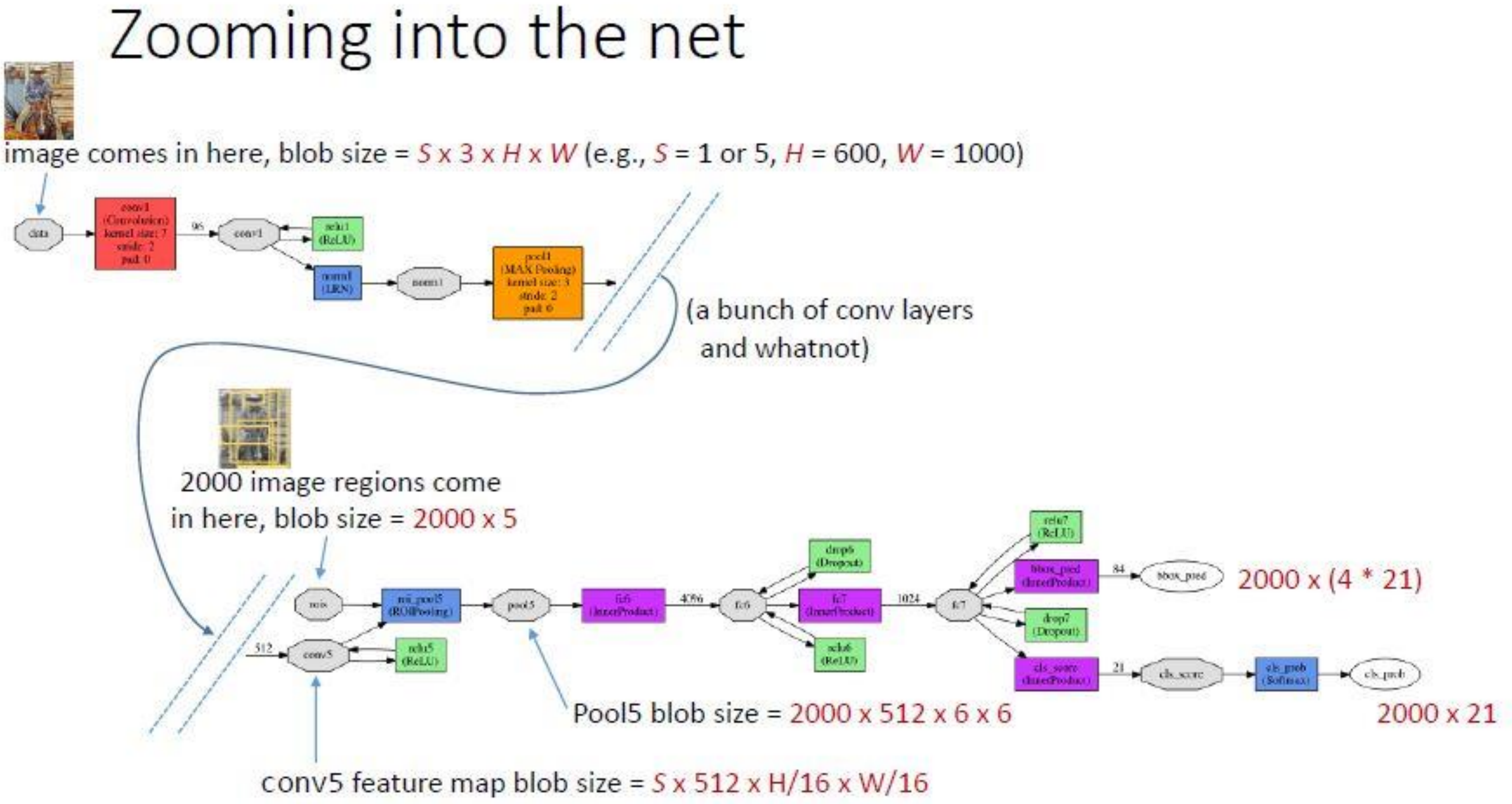


- An input image and multiple regions of interest (RoIs) are input into a fully convolutional network.
- Each RoI is pooled into a fixed-size feature map and then mapped to a feature vector by fully connected layers (FCs).
- The network has two output vectors per RoI: softmax probabilities and per-class bounding-box regression offsets.
- The architecture is trained end-to-end with a multi-task loss.

Fast R-CNN

Fast R-CNN

Ross Girshick / **ICCV2015**



Zooming into the net

image comes in here, blob size = S x 3 x H x W (e.g., S = 1 or 5, H = 600, W = 1000)

(a bunch of conv layers and whatnot)

2000 image regions come in here, blob size = 2000 x 5

2000 x (4 * 21)

Pool5 blob size = 2000 x 512 x 6 x 6

2000 x 21

conv5 feature map blob size = S x 512 x H/16 x W/16

http://tmmse.xyz/deep-learning-based-detection/

Faster R-CNN

Faster R-CNN
= Faster Region-based Convolutional Networks
for object detection

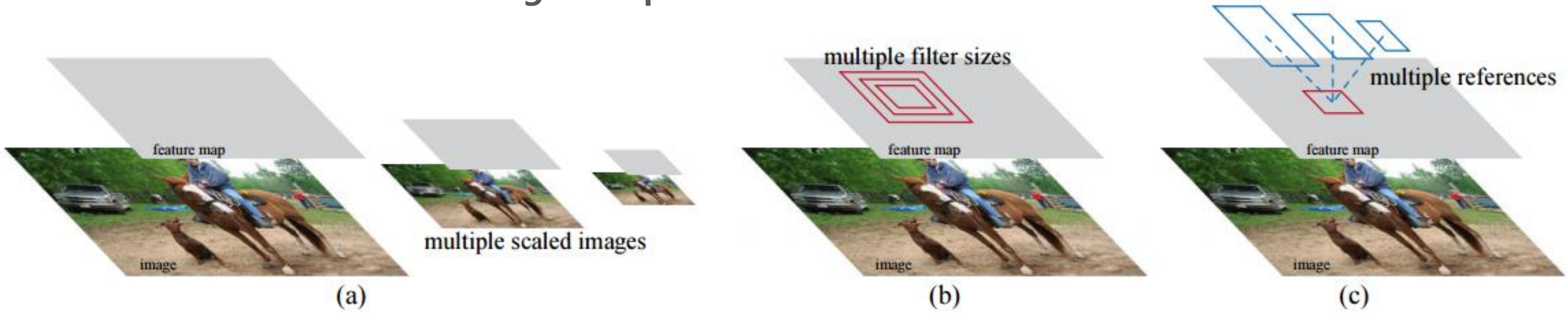
How?

Faster R-CNN

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun / **NIPS2015**

Different schemes for addressing multiple scales and sizes.



(a) Pyramids of images and feature maps are built, and the classifier is run at all scales.
(b) Pyramids of filters with multiple scales/sizes are run on the feature map.
(c) Using pyramids of reference boxes in the regression functions.

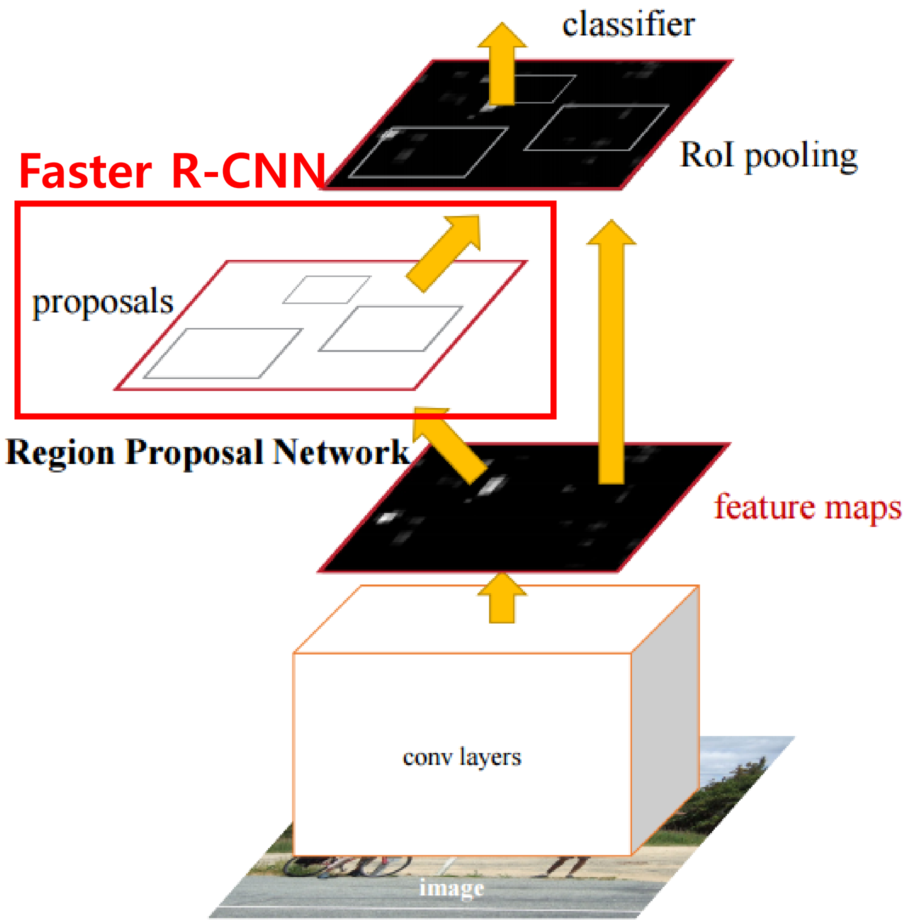
http://tmmse.xyz/deep-learning-based-detection/

Faster R-CNN

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun / **NIPS2015**

Architecture of Faster R-CNN



Faster R-CNN

- a deep fully convolutional network
- the Fast R-CNN detector

Sharing Features for RPN and Fast R-CNN

- adopt Fast R-CNN
- describe algorithms that learn network composed of RPN and Fast R-CNN with shared convolutional layers

Three ways for training networks with features shared:

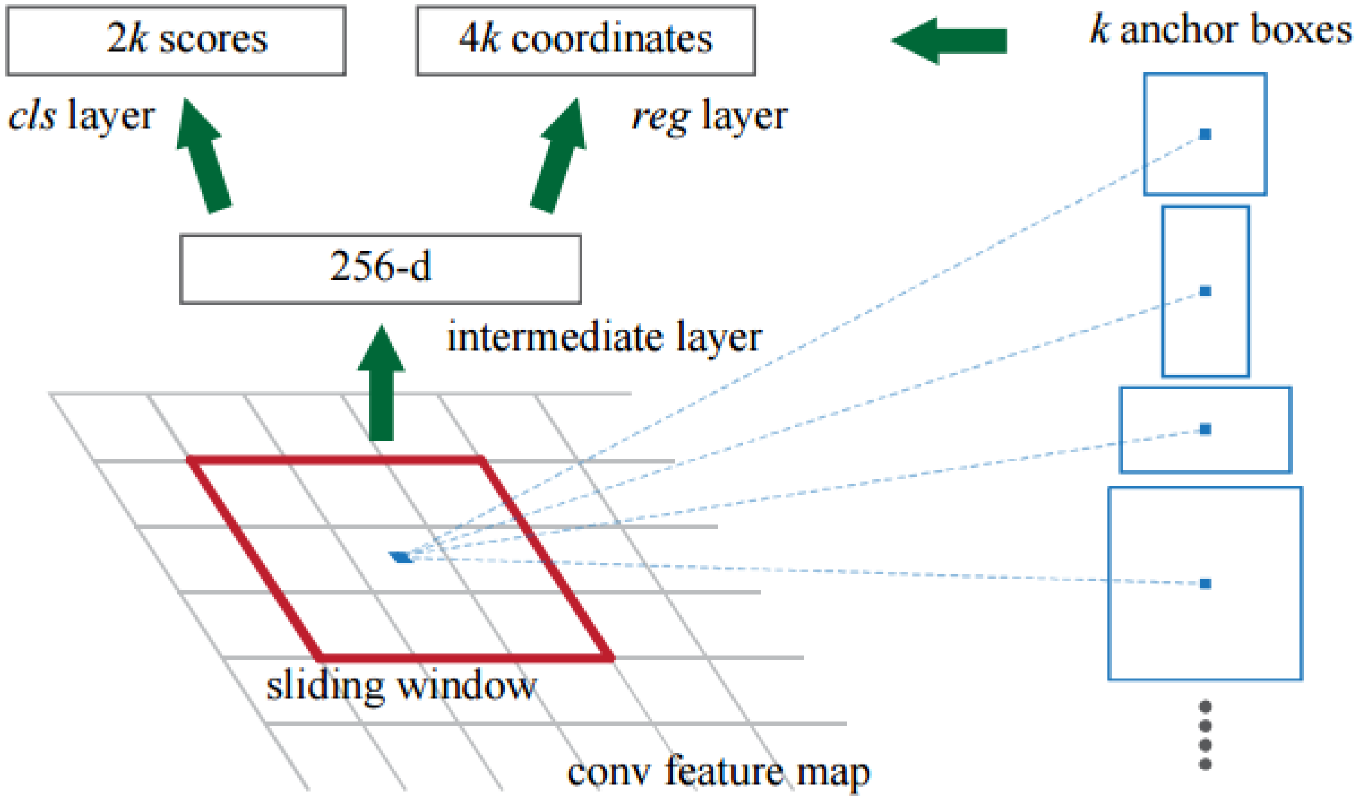
(1) Approximate joint training
(2) Non-approximate joint training

Faster R-CNN

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun / **NIPS2015**

Region Proposal Networks



Faster R-CNN

R-CNN -> SPPnet -> Fast-RCNN -> Faster-RCNN

- ~~When region proposal was classified in CNN, performance level is low due to image modify/loss.~~
- ~~Velocity is low because CNN computation was practiced after classifying region proposals.~~
- ~~Algorithm using in region proposal is not beneficial for fast calculation on GPU.~~

Fast-RCNN solve

Faster-RCNN solve

Thanks for your attention

Reference

[1] Chen, Matthew. "Pedestrian Detection with RCNN.

[2] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

[3] He, Kaiming, et al. "Spatial pyramid pooling in deep convolutional networks for visual recognition." *European Conference on Computer Vision*. Springer International Publishing, 2014.

[4] Girshick, Ross. "Fast r-cnn." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

[5] Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems*. 2015.