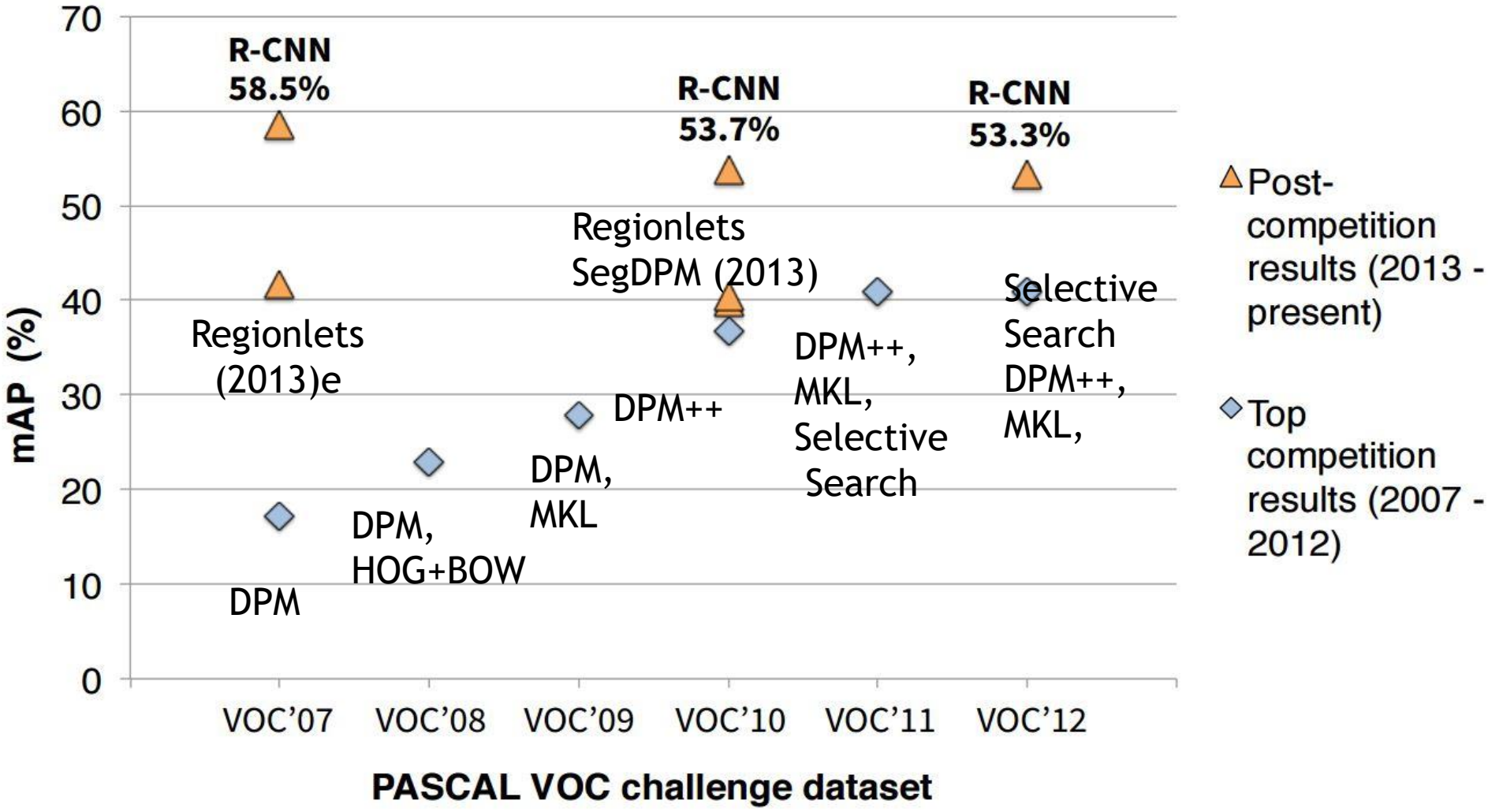
Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation

Authors: Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik

Presented by Huihuang Zheng

Problem: Object Detection

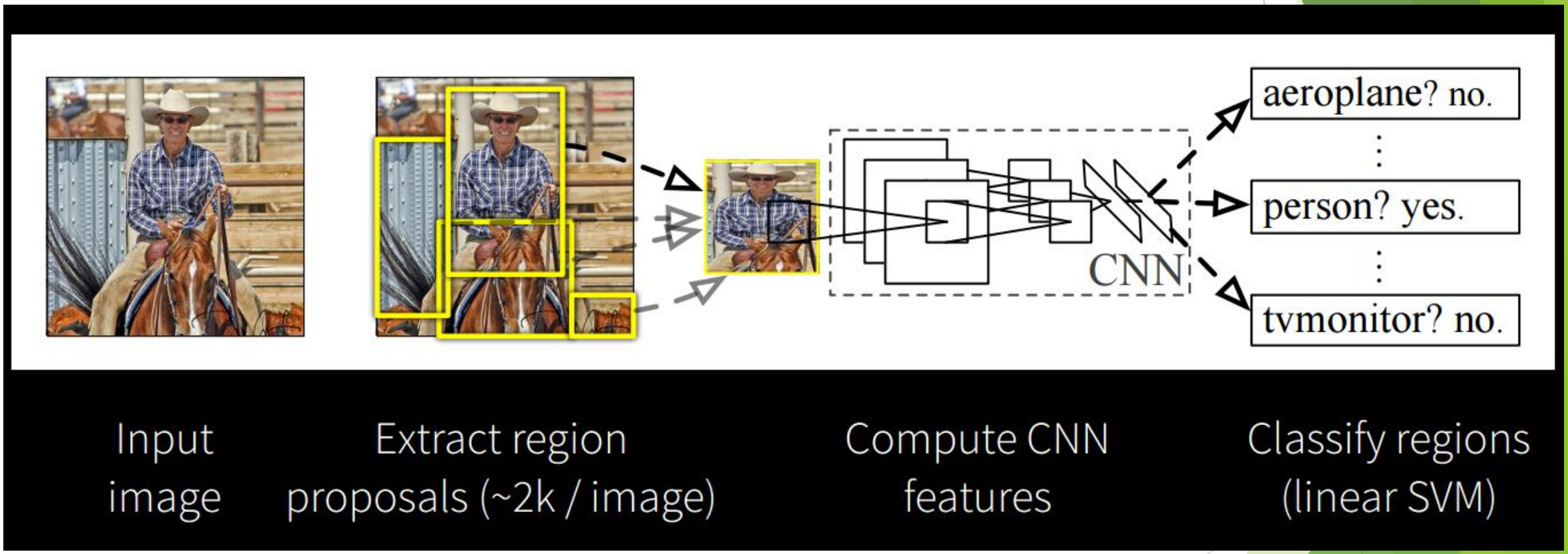


Source: http://www.cs.berkeley.edu/~rbg/slides/rcnn-cvpr14-slides.pdf

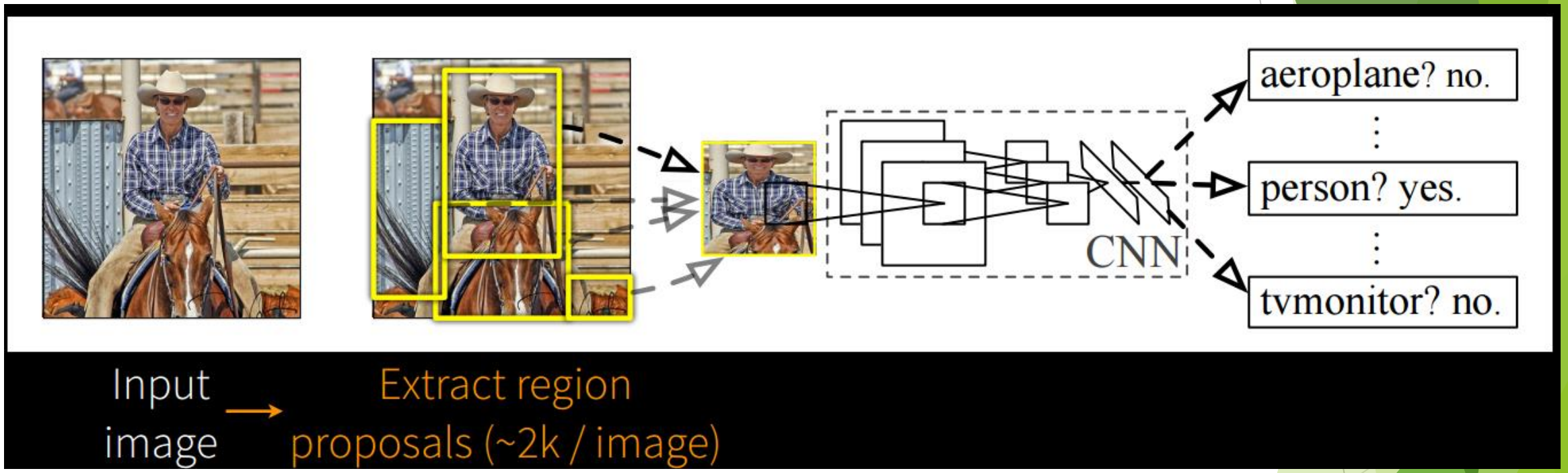
Feature Learning with CNN

- Previous best-performance methods:
  - plateaued,
  - complex
- This paper: simple, scalable
  - Two main contributions:
    - Apply CNN to bottom-up region proposals to localize
    - Fine-tune the CNN when lack of training data

Main Procedure



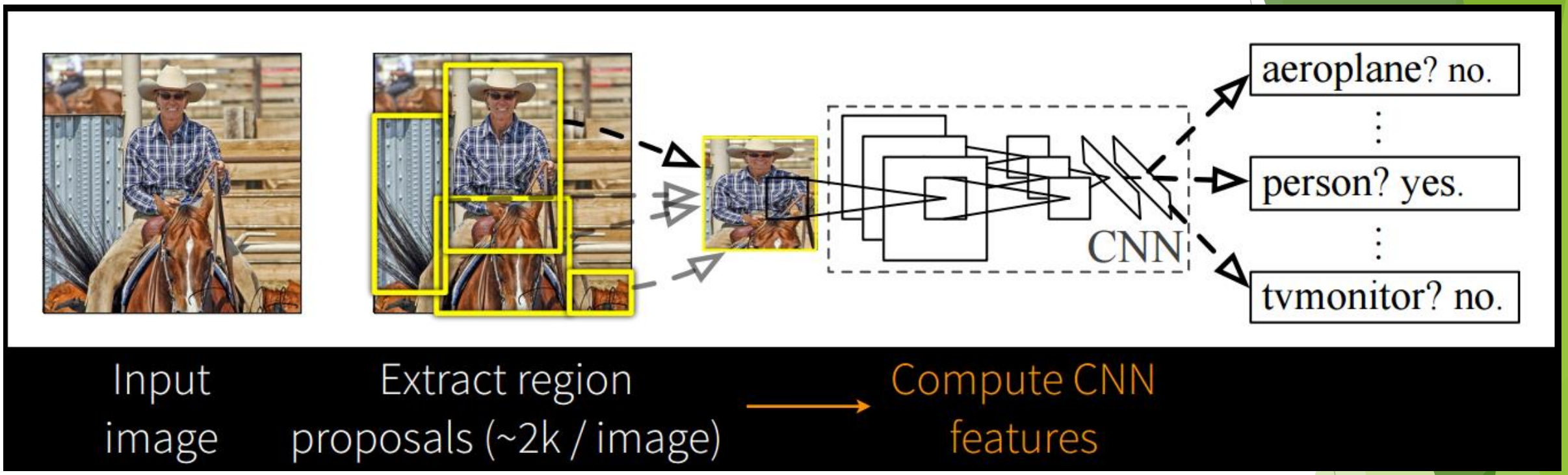
Step 1: Extract Region Proposals

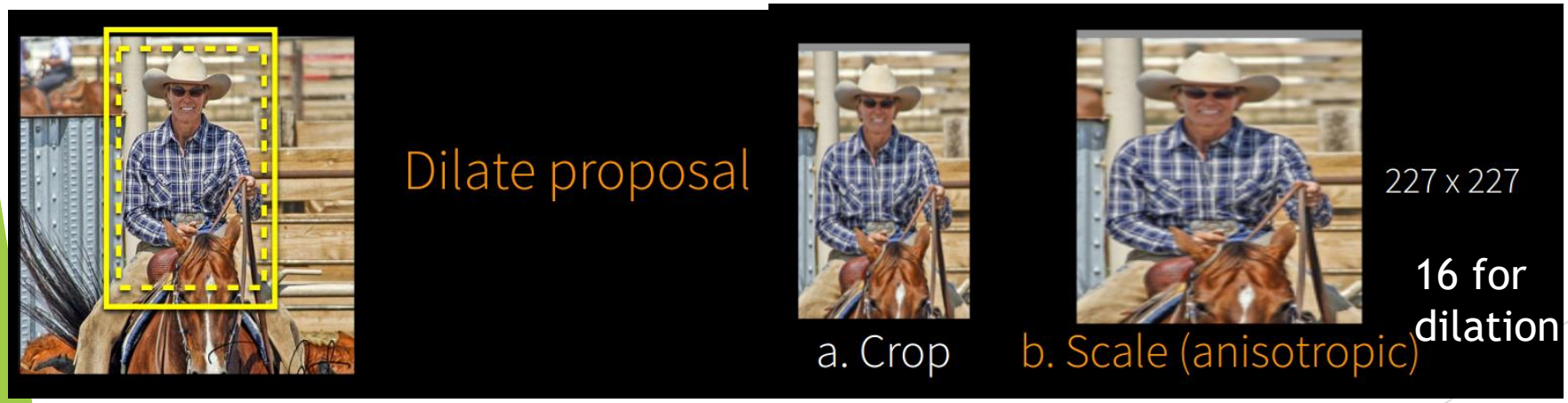


Region Proposals: many choices

- Selective Search [Uijlings et al.] (Used in this work)
- Objectness [Alexe et al.]
- CPMC [Carreira et al.]
- Category independent object proposals [Endres et al.]

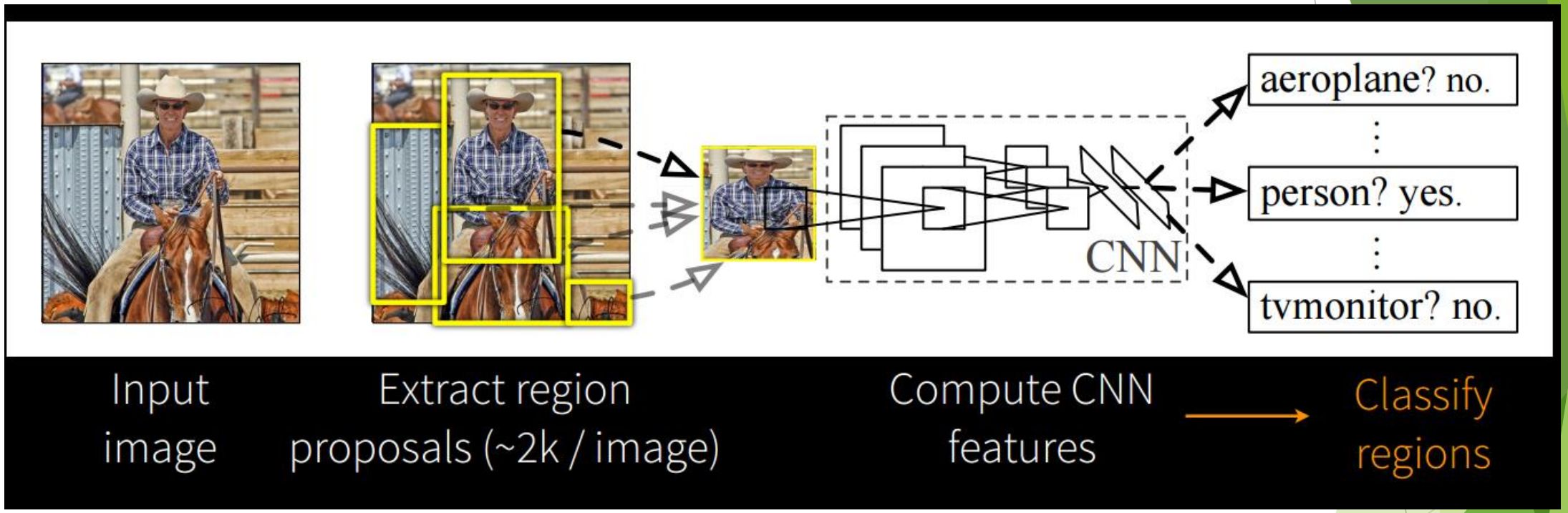
Step 2: CNN Feature





- c. Forward propagation, extract "fc7" layer feature
- Krizhevsky's AlexNet

Step 3: Classify Regions



Linear Classifier:

- SVM
  - SVM here improves accuracy! (50.9% to 54.2%) CNN classifier doesn't stress on precise location
  - SVM will be trained with hard negatives while CNN was trained with random background
- Softmax

Step 4: Modify Regions

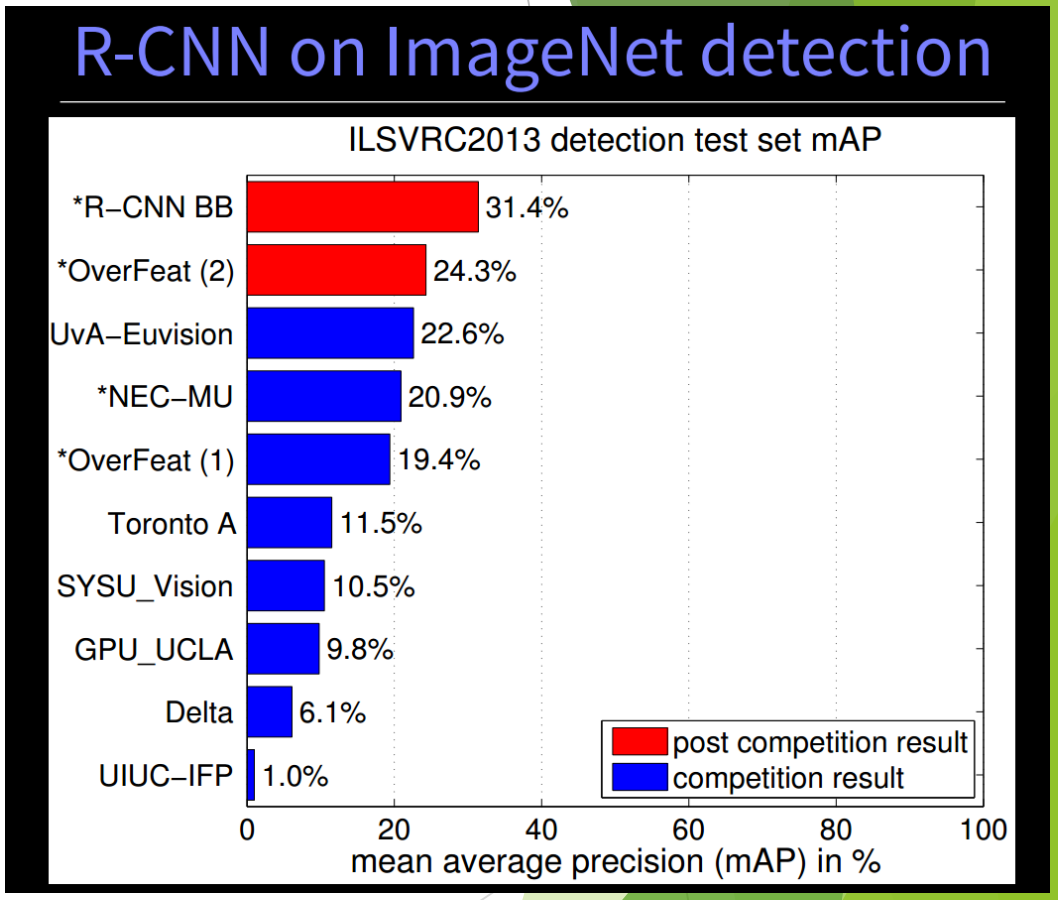
- A lot of scored regions
- Reject regions with
  - intersection-over-union (IoU) overlap with a higher scoring selected region (learned threshold)
- Bounding box regression
  - Get higher accuracy

Training: What if we lack of training data

- Solution:
  - Use pre-trained CNN (the one trained with sufficient data)
  - Fine-tune to specific task.
  - Fine-tuning also increases accuracy.
- Details in paper:
  - AlexNet [Krizhevisky et al.]
  - Stochastic gradient descent (SGD) with learning rate of 0.001, (1/10 of initial)
  - Replace 1000-way classification layer to 21-way
  - Region with >= 0.5 IoU overlap with ground-truth box as positive, others as negative.

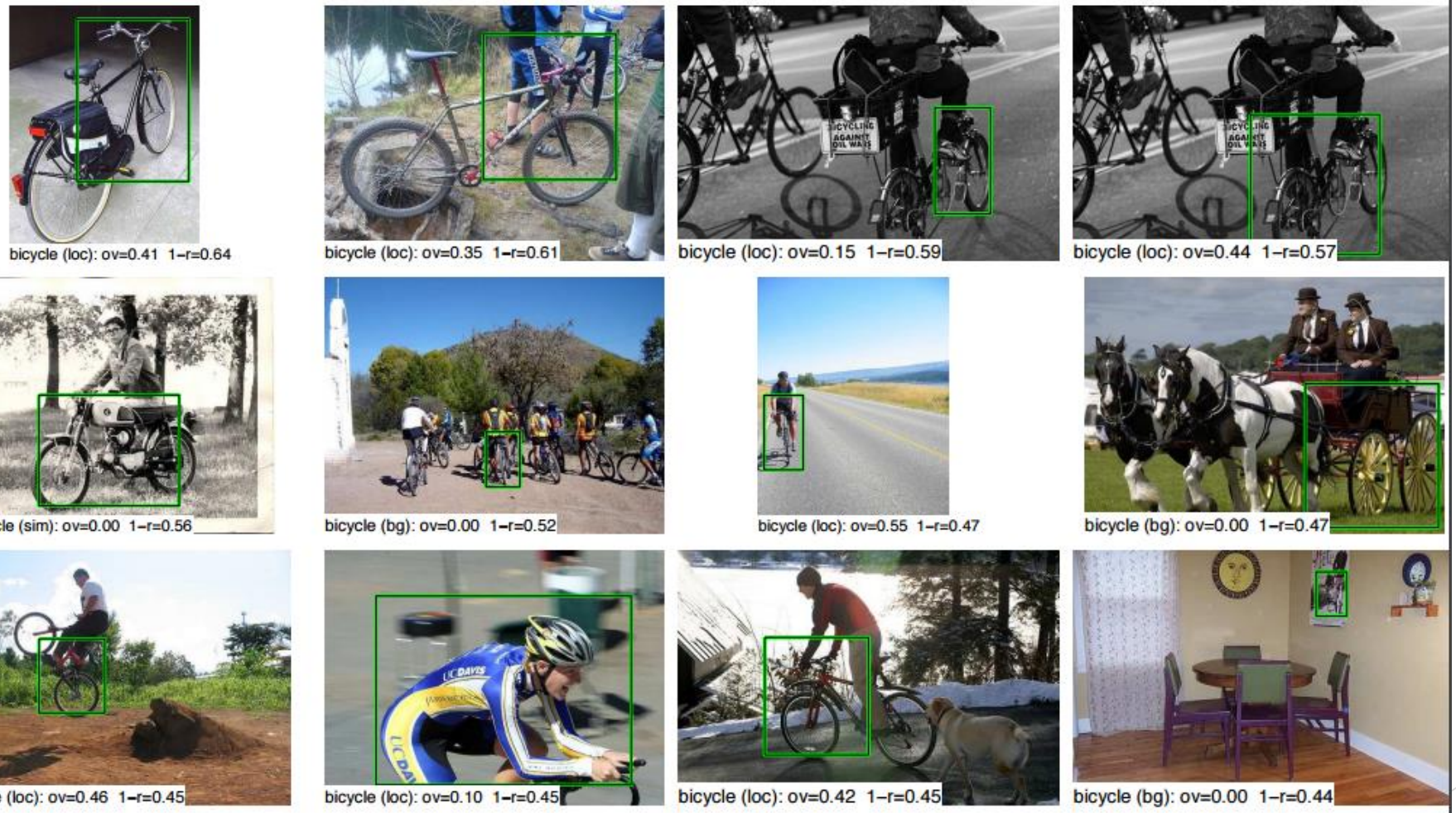
Experiment Result

	VOC 2007	VOC 2010
DPM v5 (Girshick et al. 2011)	33.7%	29.6%
UVA sel. search (Uijlings et al. 2013)		35.1%
Regionlets (Wang et al. 2013)	41.7%	39.7%
SegDPM (Fidler et al. 2013)		40.4%
R-CNN	54.2%	50.2%
R-CNN + bbox regression	**58.5%**	**53.7%**



Source: http://www.cs.berkeley.edu/~rbg/slides/rcnn-cvpr14-slides.pdf

Top bicycle FPs (AP = 72.8%)



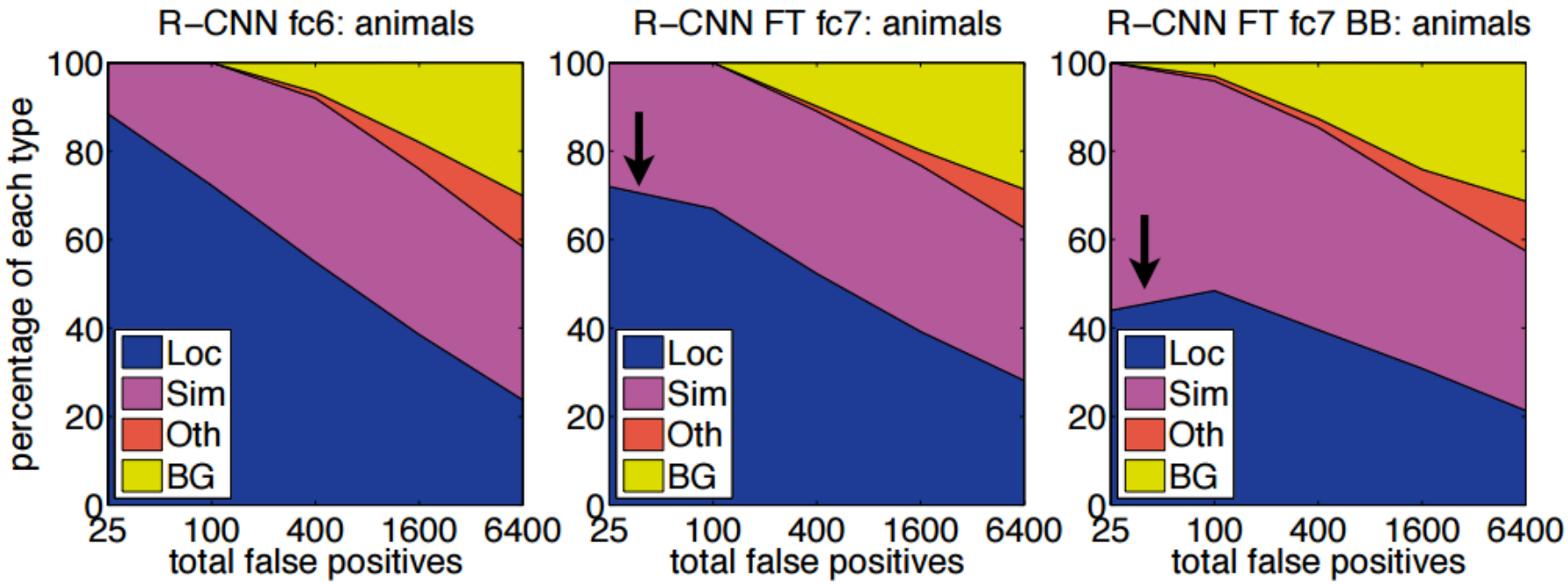
Source: http://www.cs.berkeley.edu/~rbg/slides/rcnn-cvpr14-slides.pdf

False positive #15

(zoom)

Unannotated bicycle

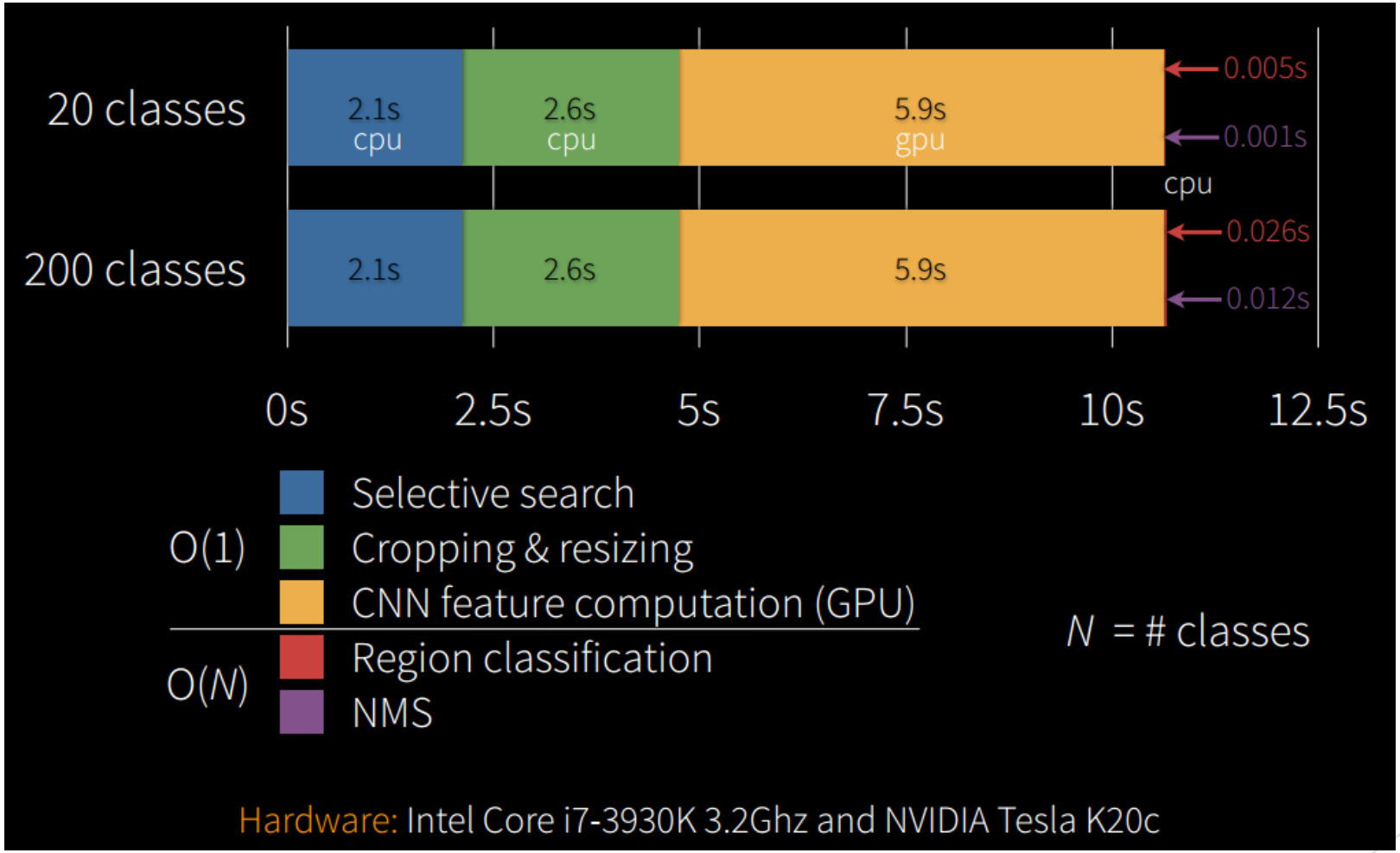
How does fine-tuning and bounding box influence result



Left: without fine-tuning, middle: with fine-tuning, right: with fine-tuning and bounding box

- Conclusion:
  - Error type of R-CNN is more about location. Suggesting that CNN feature is more discriminative
  - Bounding box helps significantly in location problem.

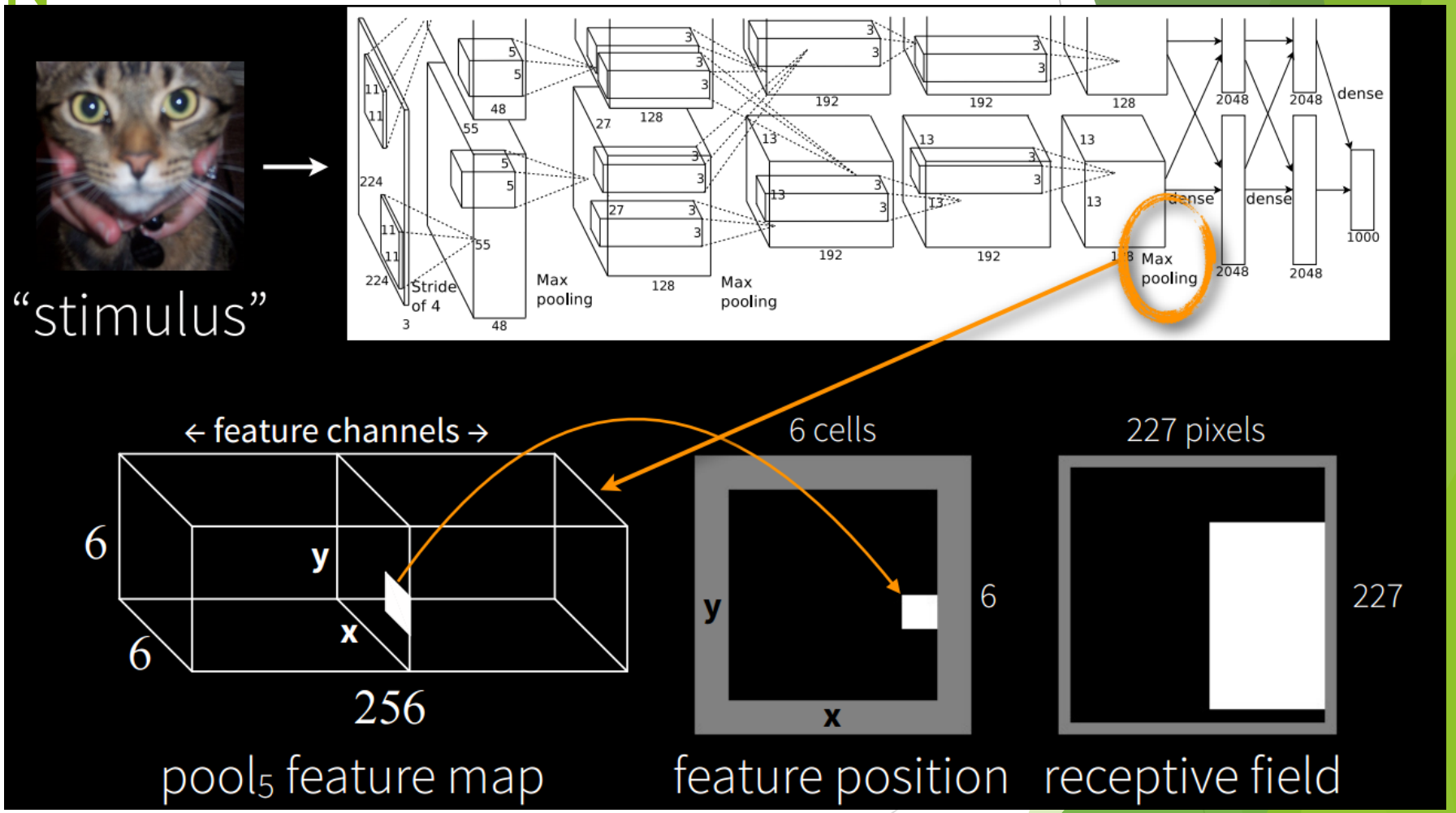
Detection Speed and Scalability



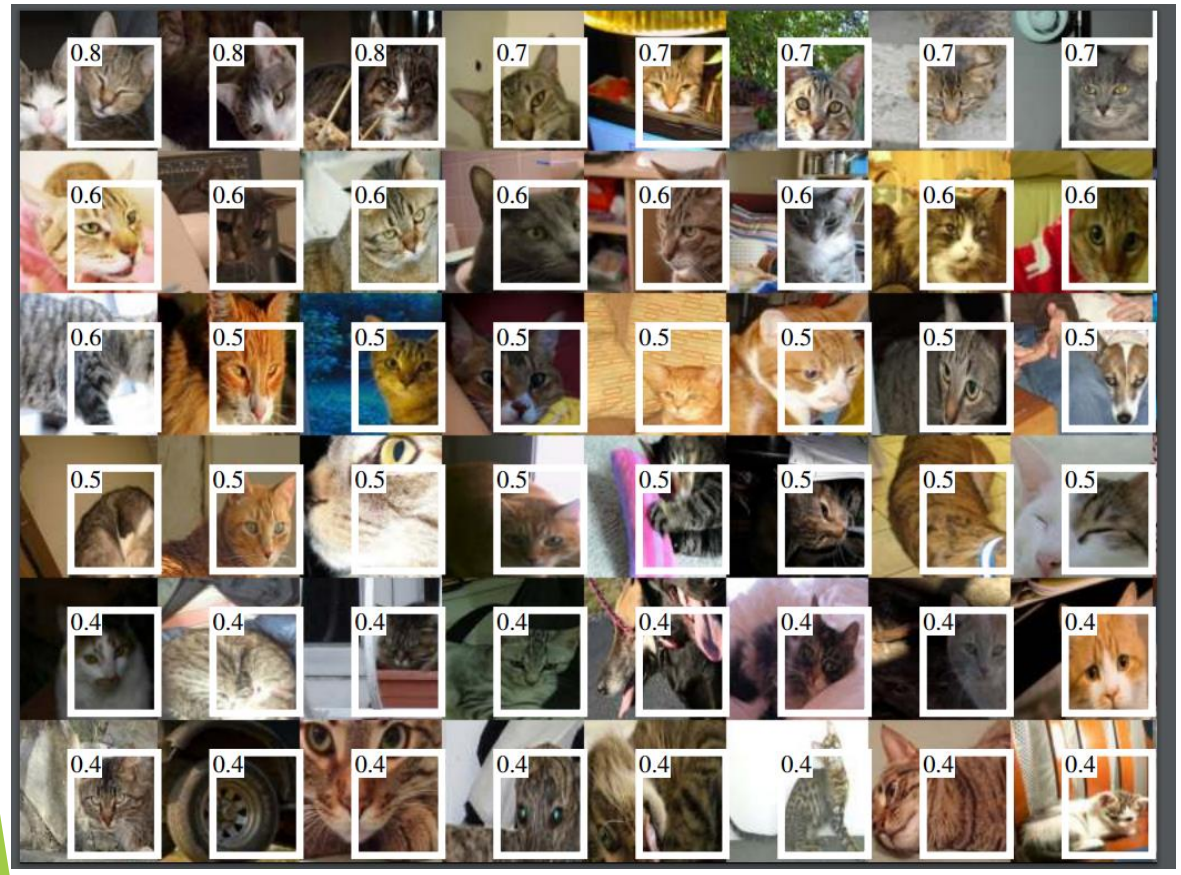
Source: http://www.cs.berkeley.edu/~rbg/slides/rcnn-cvpr14-slides.pdf

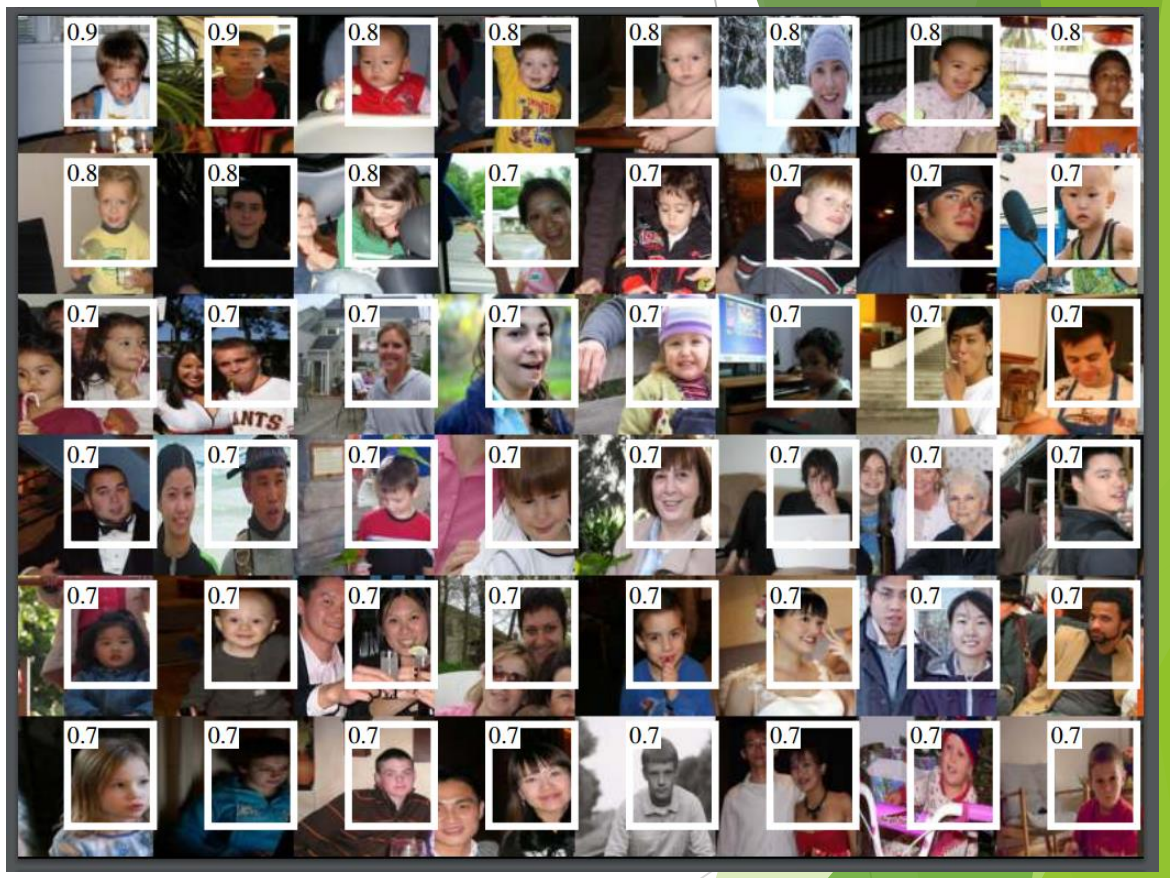
Interesting visualization: what was learnt by CNN

- Visualizing method:
  - Neurons with highest activation
  - Receptive field



Visualization: some interesting images

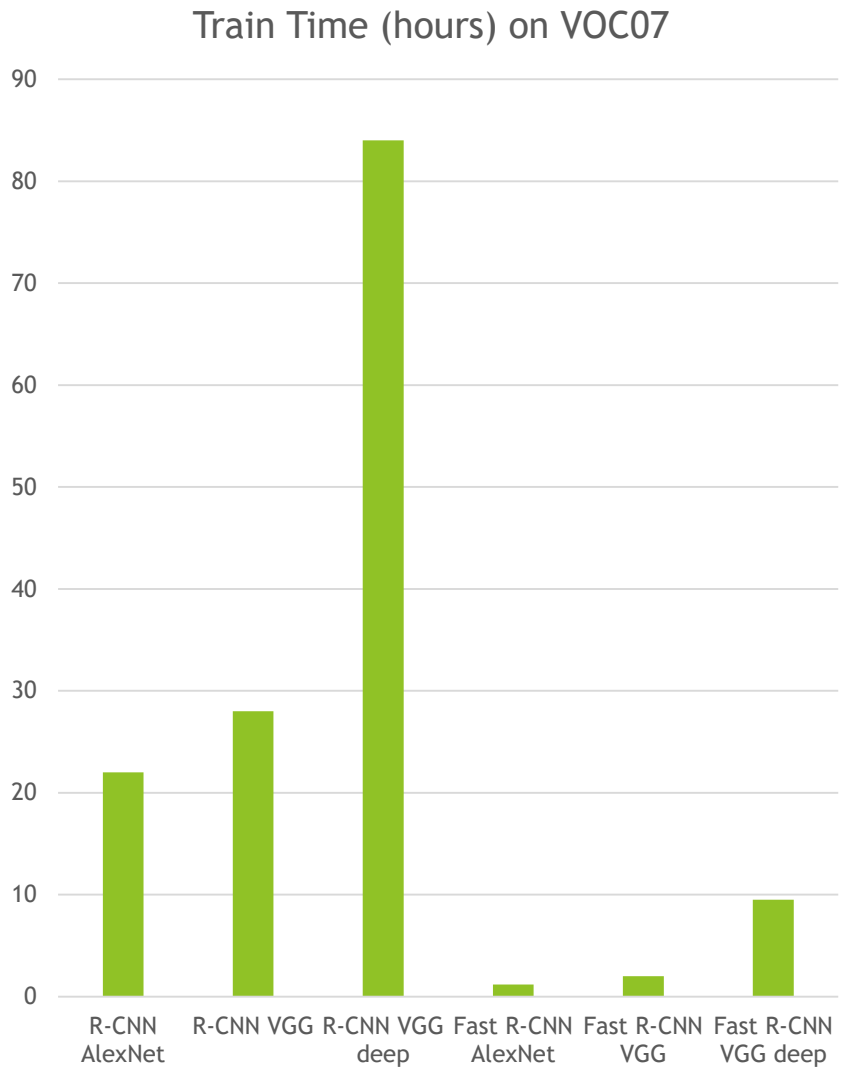


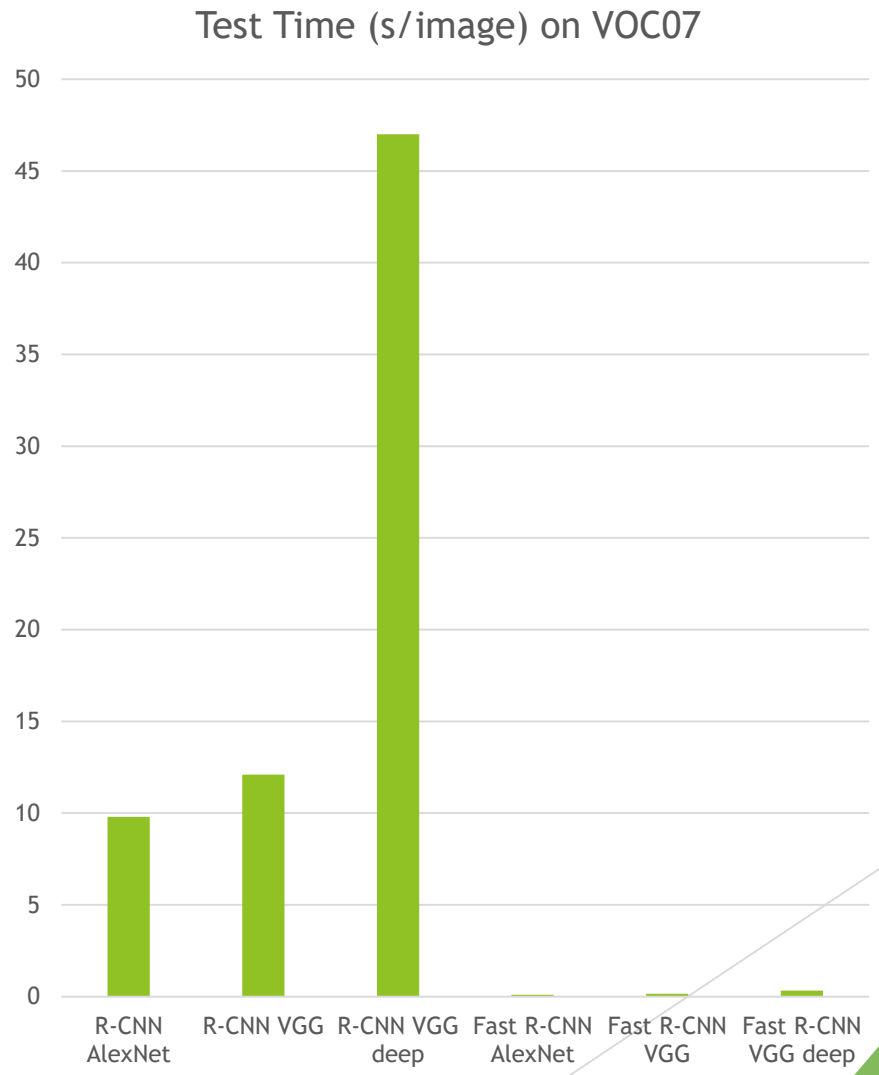


Related Future Work Papers

- Fast R-CNN, by Ross Girshick
  - R-CNN is slow, training is multi-stege, features from each object proposal
    - Sharing computation by computing a convolutional feature map for entire input image
  - Fast R-CNN Main idea: Compute a global feature map, computing region of interest in pooling layer, full-connected layer to give prediction and location.
- Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks by Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun
  - Bottleneck of Fast R-CNN is region proposals
  - Faster R-CNN computes proposals with a CNN (Region Proposal Networks (RPN))

Time Comparison





Discussion & Questions

- 1. Is simple scale the best way to make region proposals capable for CNN input?
- 2. If we have a more precise CNN, will the object detection framework in this paper be better?
- 3. Why do we use SVM at top layer?
- 4. Is fc7 better for detection and fc6 better for localization and segmentation?

- Thank you!