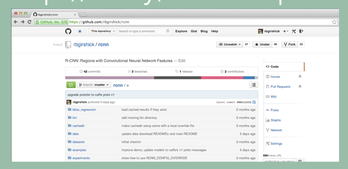
Get code & models!
http://bit.ly/rcnn-cvpr14



Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation

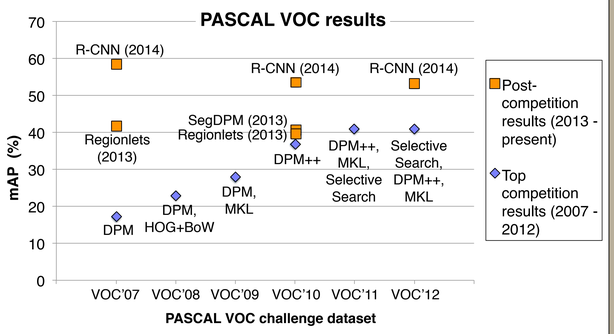
Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik

Berkeley UNIVERSITY OF CALIFORNIA

R-CNN: *Regions with CNN features*

What is R-CNN? R-CNN is a state-of-the-art object detection system that combines bottom-up region proposals with rich features learned by a convolutional neural network (CNN).

Why is R-CNN interesting? R-CNN dramatically outperforms all existing methods on PASCAL VOC and ImageNet detection. Open-source code is available for the whole system @ bit.ly/rcnn-cvpr14.

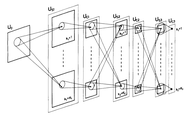


Feature learning with CNNs

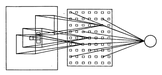
SIFT, HOG, LBP, etc.: All previous PASCAL detection top-performers were based on these low-level features.

Feature learning: R-CNN is the first CNN-based feature learning method to achieve state-of-the-art object detection performance on PASCAL VOC.

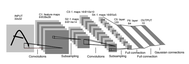
A brief history of convolutional neural networks



Fukushima 1980
Neocognitron



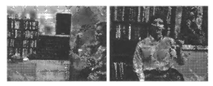
Rumelhart, Hinton, Williams
1986 "T" vs. "C" problem



LeCun et al. 1989 - 1998
Handwritten digit reading

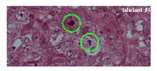
CNNs for object detection

Vaillant, Monrocq, LeCun 1994
Multi-scale face detection



Nowlan & Platt 1995
Hand tracking

LeCun, Huang, Bottou 2004
NORB dataset

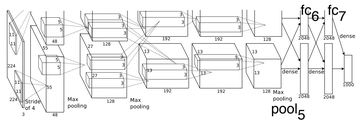


Cireşan et al. 2013
Mitosis detection

Sermanet et al. 2013
Pedestrian detection

Szegedy, Toshev, Erhan 2013
DetectorNet (PASCAL mAP 30.5%)

SuperVision CNN - Krizhevsky, Sutskever, Hinton 2012

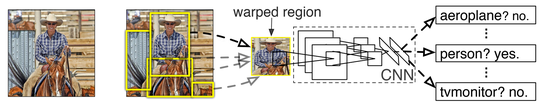


- **Breakthrough ImageNet classification results:**

ILSVRC 2012 test	Top-5 error
Fisher Vectors (ISI)	26.2%
5 SuperVision CNNs	16.4%
7 SuperVision CNNs	**15.3%**

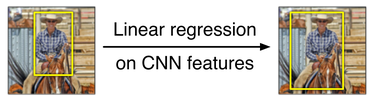
None of these methods outperform HOG-based approaches on PASCAL VOC detection

R-CNN at test time



1. Input image **2.** Extract region proposals (~2k) **3.** Compute CNN features **4.** Classify regions

Proposal refinement by regression



Original proposal — Predicted object bounding box

- For each proposal, predict object bounds
- Transfer scores from original proposal
- Trained with regularized least squares

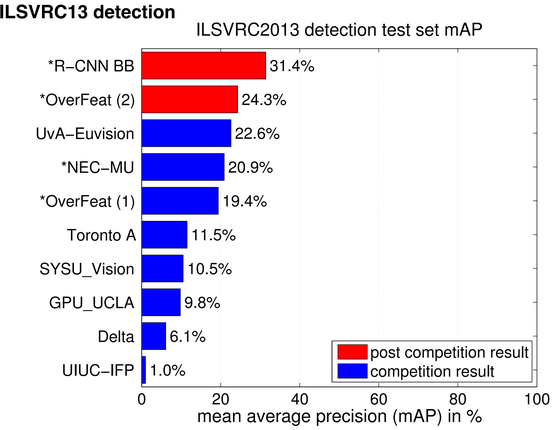
Region proposal mechanisms

- **Selective search** (Uijlings et al. 2013) used in this work
- Many other choices: **objectness** (Alexe et al.), **category independent object proposals** (Endres & Hoiem), **CPMC** (Carreira & Sminchisescu), **MCG** (Abeláez et al.), **BING** (Ming et al.), …

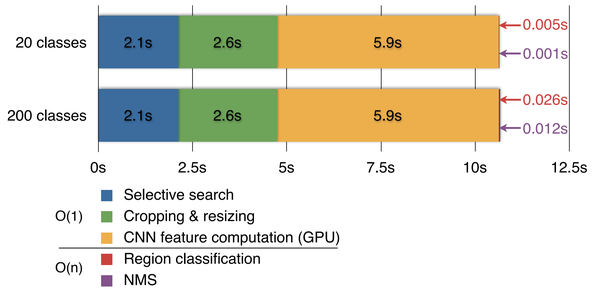
PASCAL VOC test results	VOC 2007	VOC 2010
DPM v5 (Girshick et al. 2011)	33.7%	29.6%
UVA sel. search (Uijlings et al. 2013)	–	35.1%
Regionlets (Wang et al. 2013)	41.7%	39.7%
SegDPM (Fidler et al. 2013)	–	40.4%
R-CNN	54.2%	50.2%
R-CNN + bbox regression	**58.5%**	**53.7%**

Metric: mean average precision (higher is better)

ILSVRC13 detection



Detection speed and scalability



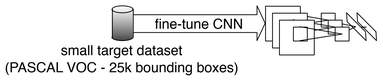
Hardware: Intel Core i7-3930K 3.2Ghz and NVIDIA Tesla K20c
We thank NVIDIA for generous hardware donations.

Training R-CNN

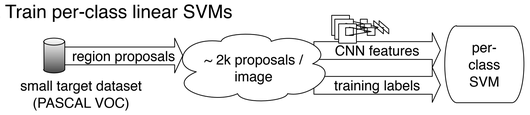
- Detection data (bounding-box labels) are scarce!
- **Key insight:** use *supervised* pre-training on a data-rich, auxiliary task

1. Pre-train CNN for **image classification**

2. Fine-tune CNN for PASCAL **detection**



3. Train per-class linear SVMs



Try Caffe! For CNN training, fine-tuning, and feature extraction we use the Caffe CNN library. http://caffe.berkeleyvision.org. Jia, 2013.

Training ablation study

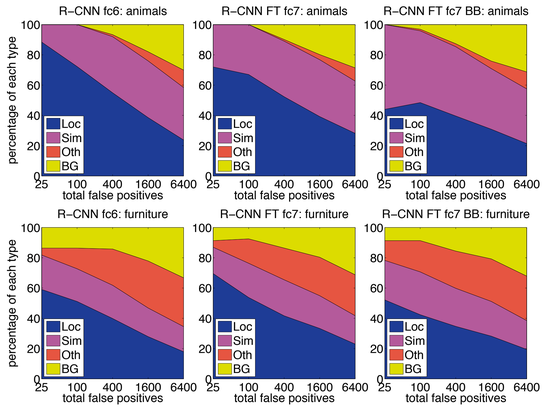
- How important is fine-tuning?
- How do different layers of the CNN perform?
- What if we use softmax instead of SVM? (mAP drops 54% to 51%)

	VOC 2007	VOC 2010
Regionlets (Wang et al. 2013)	41.7%	39.7%
SegDPM (Fidler et al. 2013)	–	40.4%
R-CNN $pool_5$ (6% of parameters)	44.2%	–
R-CNN fc_6 (71% of parameters)	46.2%	–
R-CNN fc_7 (100% of parameters)	44.7%	–
R-CNN FT $pool_5$	47.3%	–
R-CNN FT fc_6	53.1%	–
R-CNN FT fc_7	**54.2%**	**50.2%**

Metric: mean average precision (higher is better)

False positive analysis

Tools from: Hoiem, Chodpathumwan, Dai. 2012.



Semantic segmentation

- R-CNN operates on region proposals
- Extension from detection to semantic segmentation is easy

CPMC segments
(Carreira & Sminchisescu)

Compute CNN features on both "full" and "fg"

	VOC 2011
UCB Regions and Parts (Arbeláez et al.)	40.8%
Bonn second-order pooling (Carreira et al.)	**47.6%**
R-CNN fc_6 full+fg (no fine-tuning)	**47.9%**

Metric: mean segmentation accuracy (higher is better)

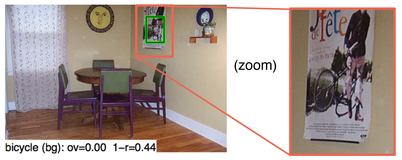
Feature visualizations

- Mixture of "grandmother cells" and distributed codes

Conclusions

- R-CNN dramatically outperforms all previous detection methods on PASCAL VOC
- R-CNN outperforms recent CNN-based detection methods, such as OverFeat, on ImageNet
- R-CNN is relatively simple and completely open source
- Detection speed is a manageable 10s / image
- Excellent scalability with number of classes
- Try R-CNN today! Get code & models at http://bit.ly/rcnn-cvpr14

A particularly pleasing detection (that was marked as an error)



Bicycle false positive #15 — Aha! It's an unannotated bicycle!

1949 movie by Jacques Tati

Look for our papers building on R-CNN

- Training R-CNN with weak supervision. Song et al. ICML'14
- Analysis of CNN training and features. Agrawal et al. ECCV'14
- 50.5% PASCAL segmentation accuracy. Hariharan et al. ECCV'14
- R-CNN with parts for fine-grained recognition. Zhang et al. ECCV'14
- Detection and segmentation in RGB-D images. Gupta et al. ECCV'14