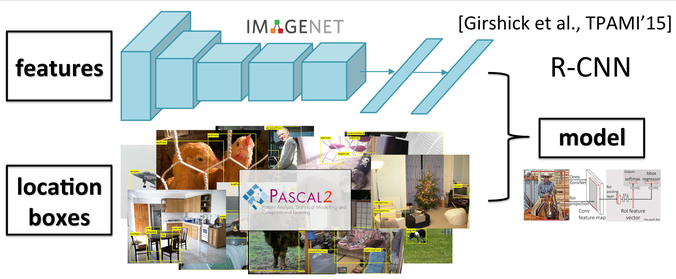
Webly Supervised Training of Convolutional Networks

Xinlei Chen Abhinav Gupta

Goal: directly train object detectors from www images

MS COCO, ~90 categories, ~2M instances — amazon mechanical turk — more categories

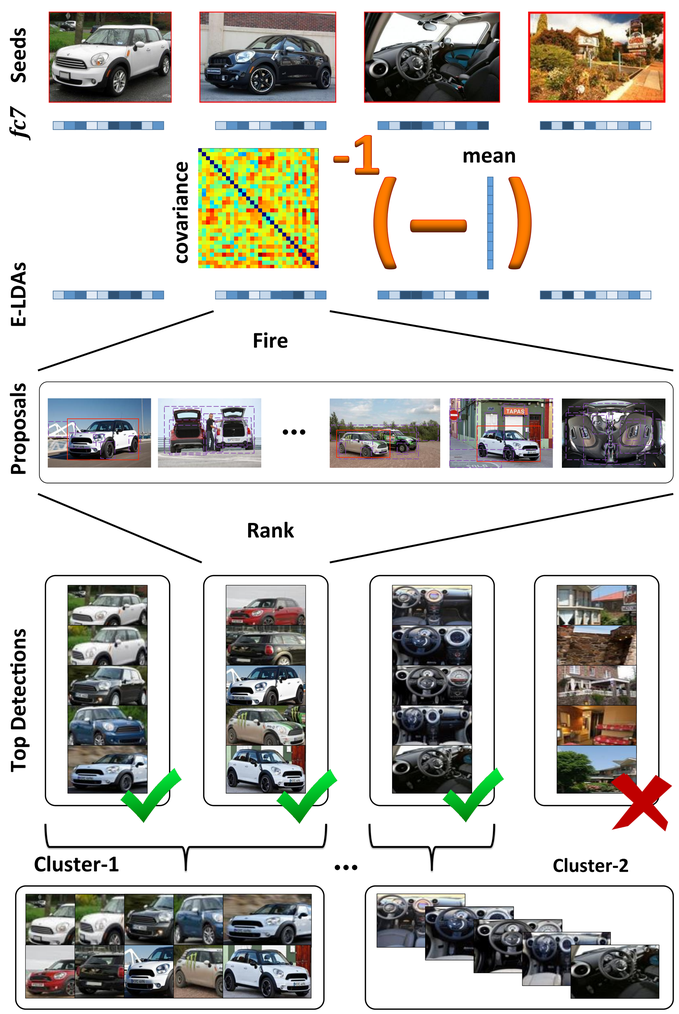


[Girshick et al., TPAMI'15]

R-CNN

Localization

[Hariharan et al., ECCV'12]



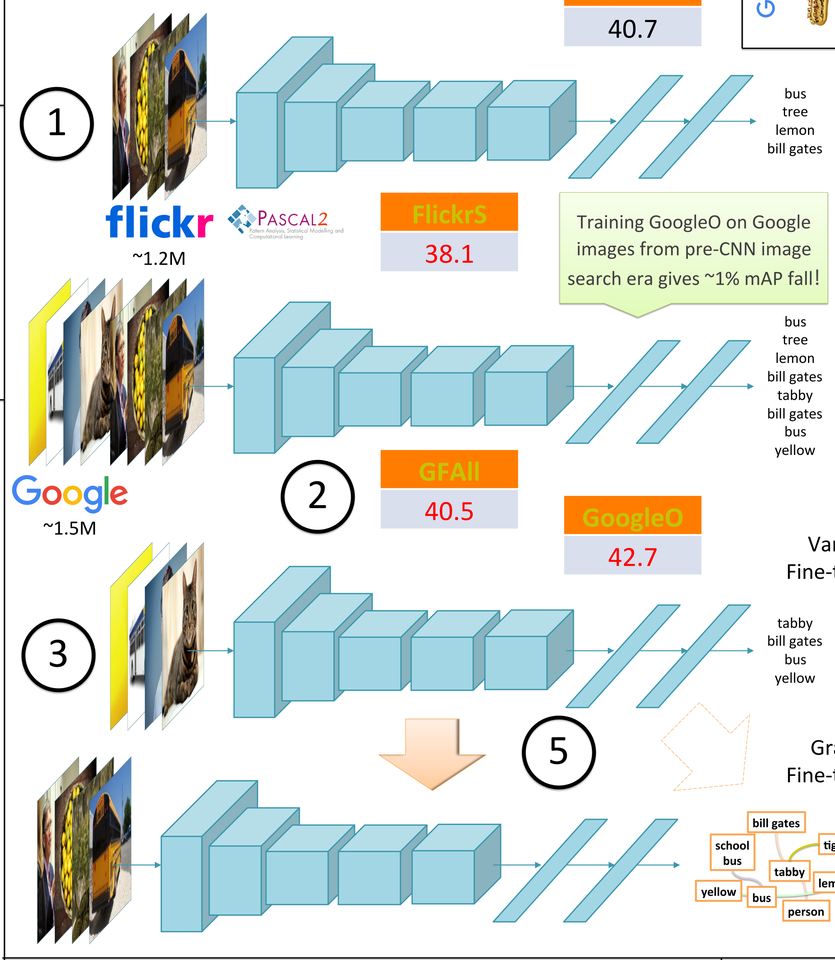
Learning Features

[krizhevsky et al., NIPS'12]
[Chen et al., ICCV'13]

- Classes: ImageNet (1K) + NEIL (1.3K), Object/Attribute
- Query, Retrieve, Predict as (Noisy) Label
- AlexNet, FC7, mean Average Precision
- R-CNN on VOC 2007 w/o Fine-tuning

ImageNet
44.7

Scratch
40.7



FlickrS
38.1

Training GoogleO on Google images from pre-CNN image search era gives ~1% mAP fall!

GFAll
40.5

GoogleO
42.7

Vanilla Fine-tuning

FlickrF
43.4

Graph Fine-tuning

$$L = \sum_k \sum_i R_{il_k} \log(CNN(i|I_k))$$

FlickrG
44.7

Observation

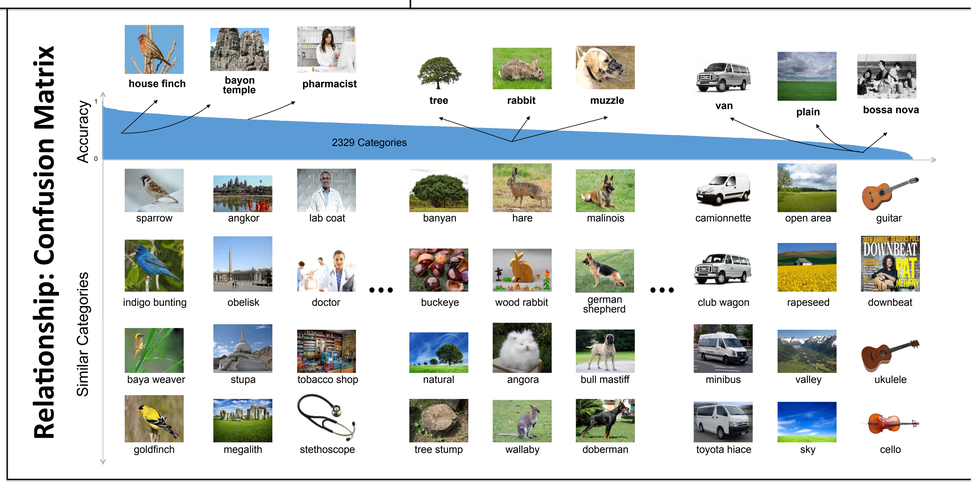
Scene Classification MIT-Indoor 67

ImageNet	Overfeat
56.8	58.4
GoogleO	FlickrG
58.1	59.2

GoogleA*
66.5

*trained with additional ~0.6M scene images

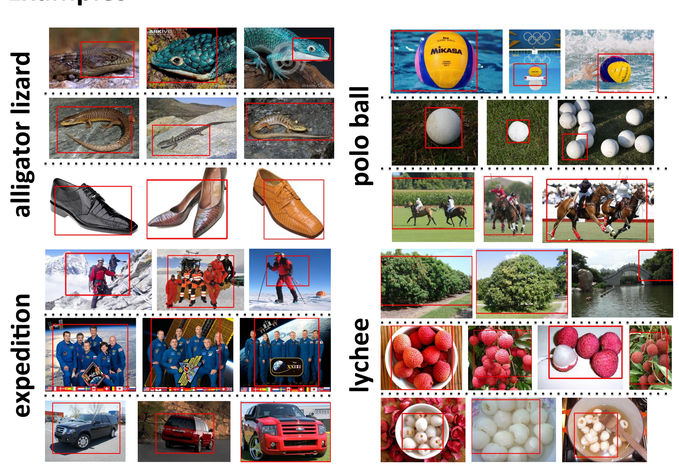
[Sermanet et al., Arxiv'13]



R-CNN With VOC Fine-tuning

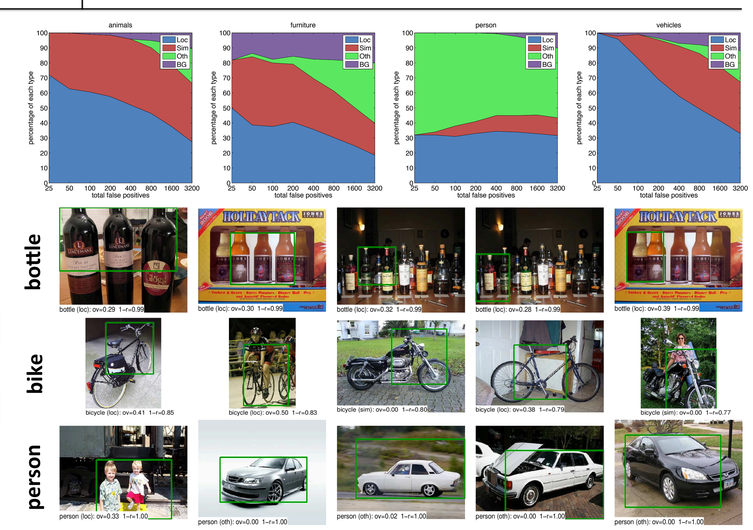
Year	ImageNet	GoogleO	FlickrG
2007	54.2	53.1	53.3
2012	52.3	52.4	52.7

Examples



Results on VOC 2007

- Basic, using all boxes in a class (FlickrG)
- EdgeBox Augmentation, IoU >= 0.5 (EA)
- Category Expansion, add similar categories (CE)

LEVAN	FlickrG
17.1	22.9
FlickrG-EA	FlickrG-CE
23.0	24.4

[Zitnick & Dollar, ECCV'14]
[Divvala et al., CVPR'14]