Object Detection Using CNNs

C.-C. Jay Kuo
University of Southern California

History in Object Detection

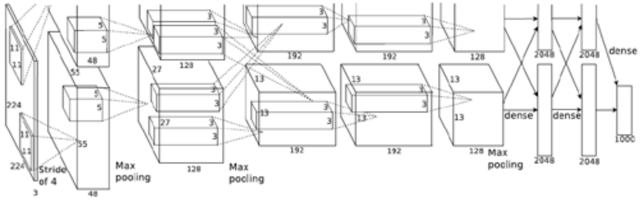
Year	Methods	Performance
1980s	Haar LBP …	~1mAP
1998	SIFT	~10mAP
2001	Cascade	~20mAP
2005	BoW HOG	~25mAP
2008	DPM	~40mAP
2013	AlexNet	~60mAP
2014	RCNN	
2015	SPP-NET FRCNN Multi-RCNN	~70mAP

mAP: Mean Average Precision

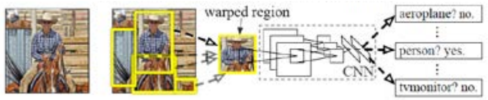
Object Detection with DNN

2013 | 2014 | 2015

AlexNet

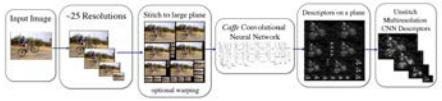


RCNN

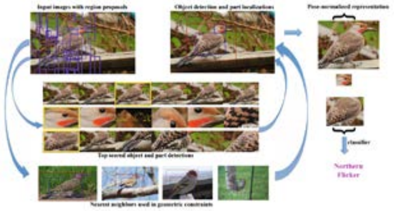


Caffe

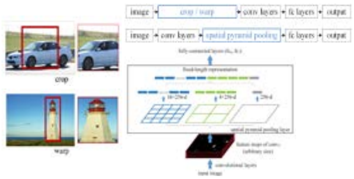
DenseNet



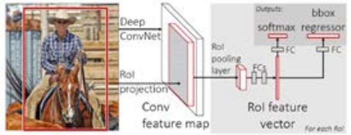
Part-Based R-CNN



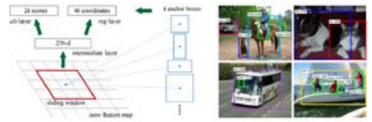
SPP-Net



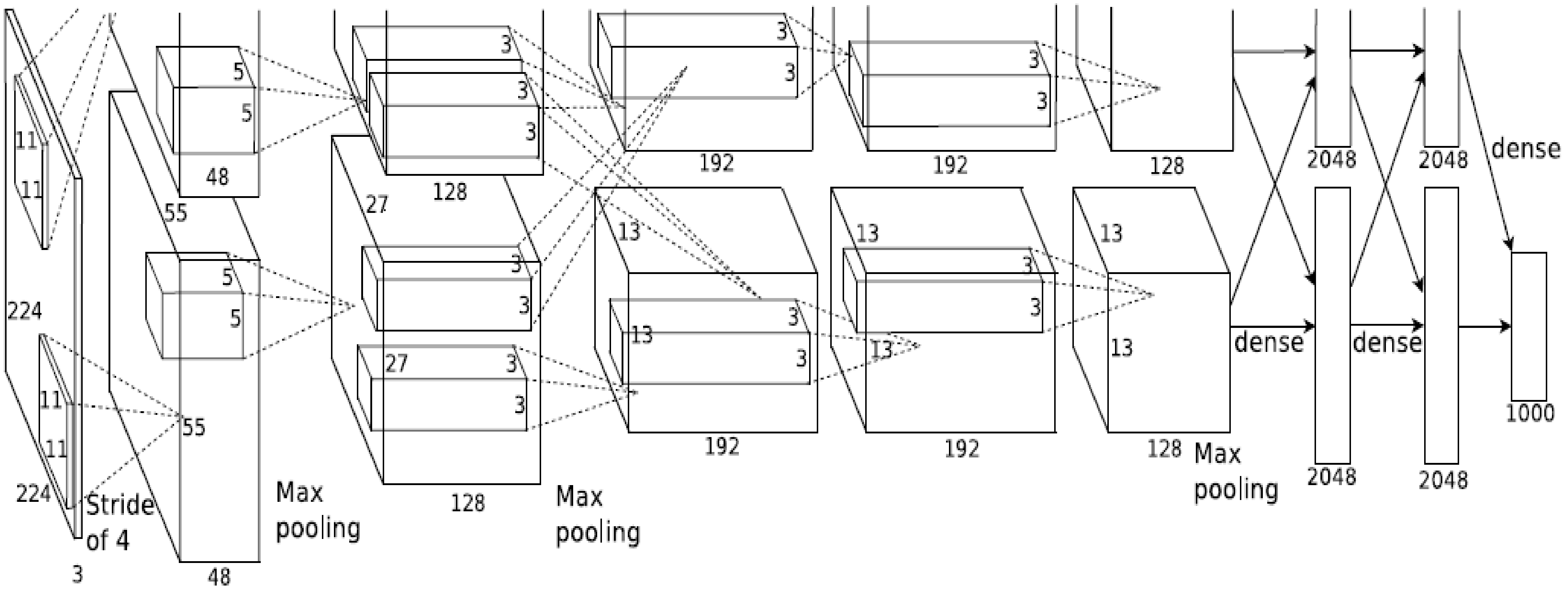
Fast R-CNN



Faster R-CNN



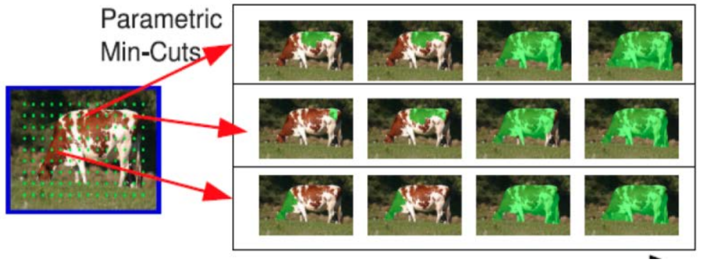
Alex Net (2013)

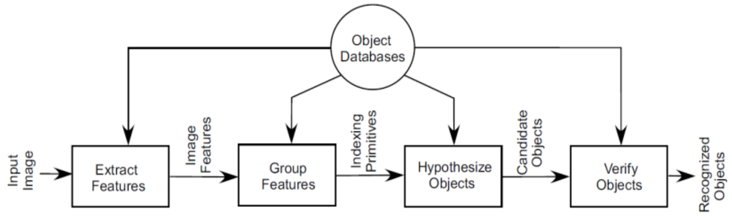


A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks.

Region Proposal with CNN (RCNN)

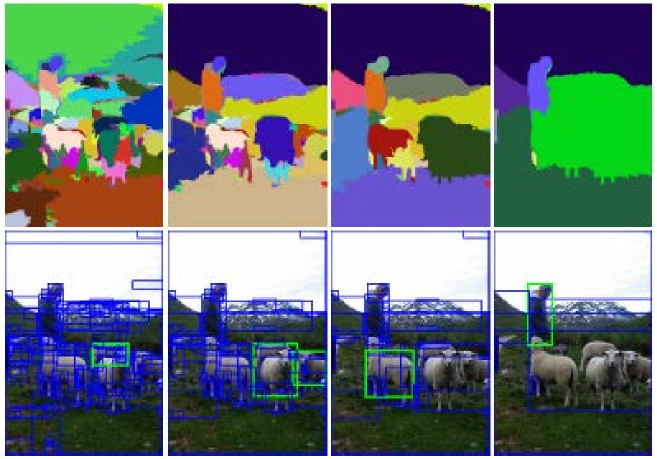
Constraint Parametric Min-Cut





J. Carreira and C.sminchisescu. CPMC Automatic object segmentation using constrained parametric min-cust. In CVPR, 2012

Selective Search

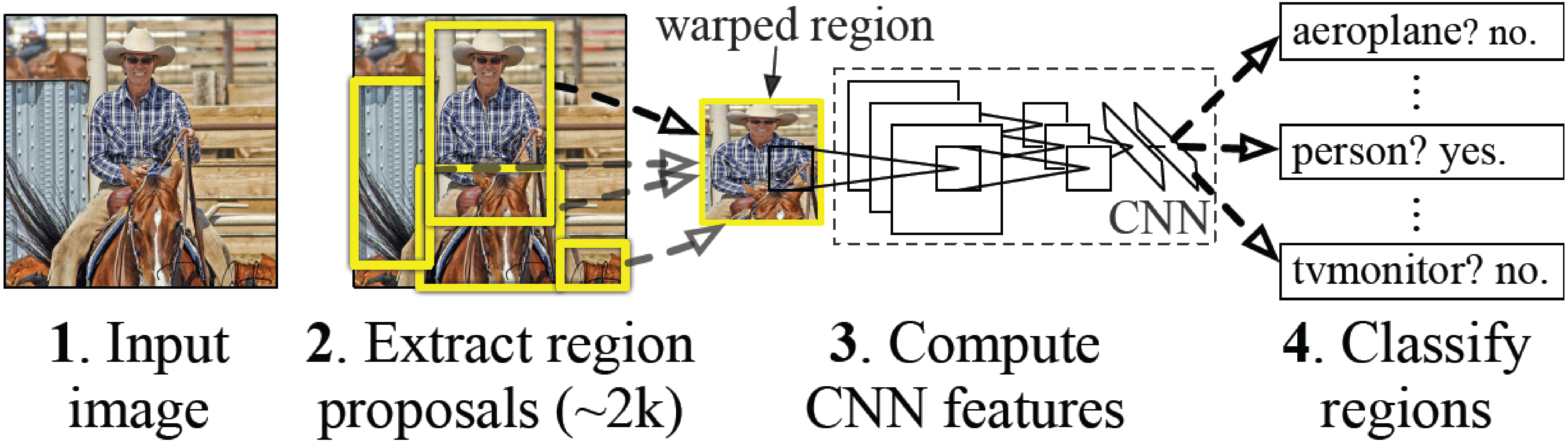


J.R.R Uijlings, K.E.A. Van De Sande, T. Gevers, and A.W.M. Smeulders, Selective Search for Object Recognition. In IJCV 2013

EdgeBox

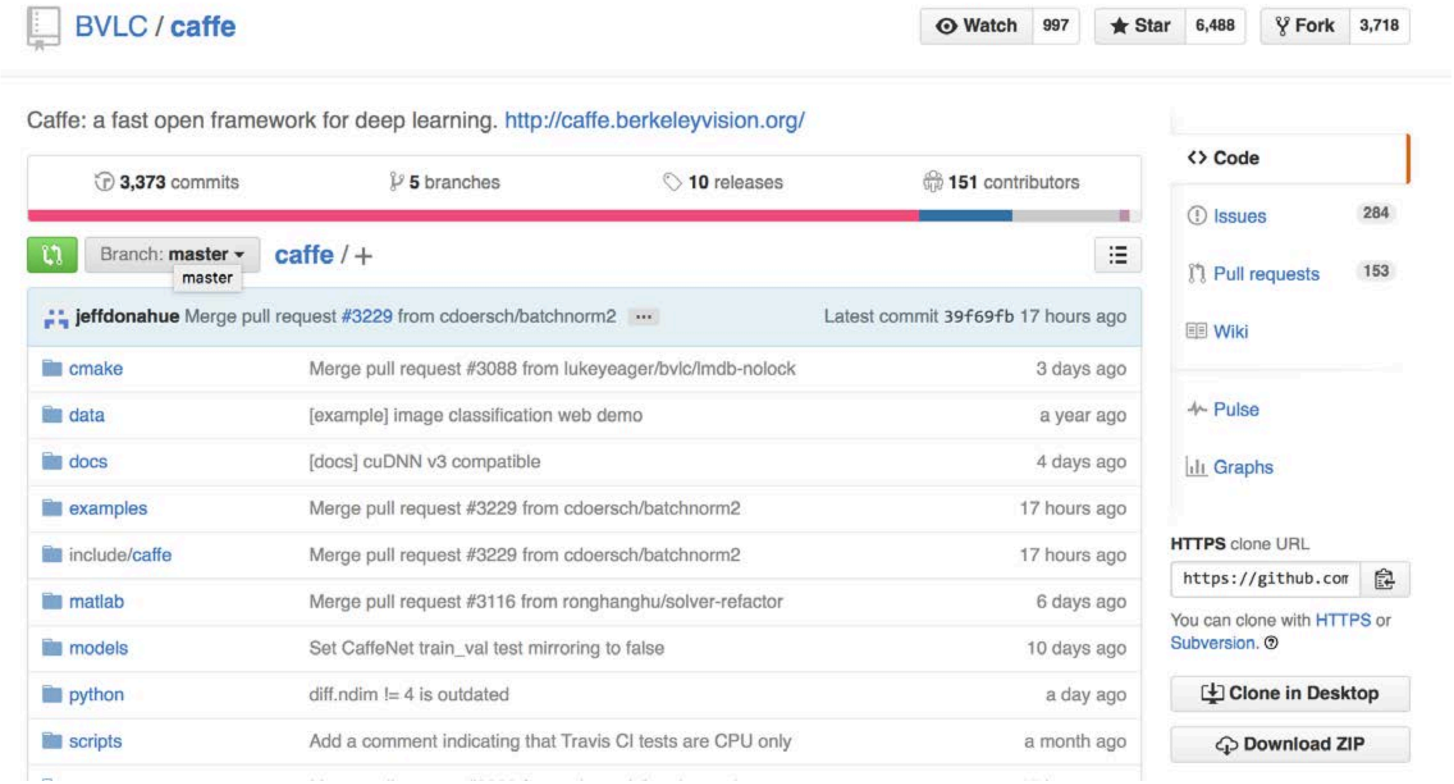
C. L. Zitnick and P. Dollar, “Edge boxes: Locating object ´ proposals from edges,” in ECCV, 2014.

RCNN Flowchart

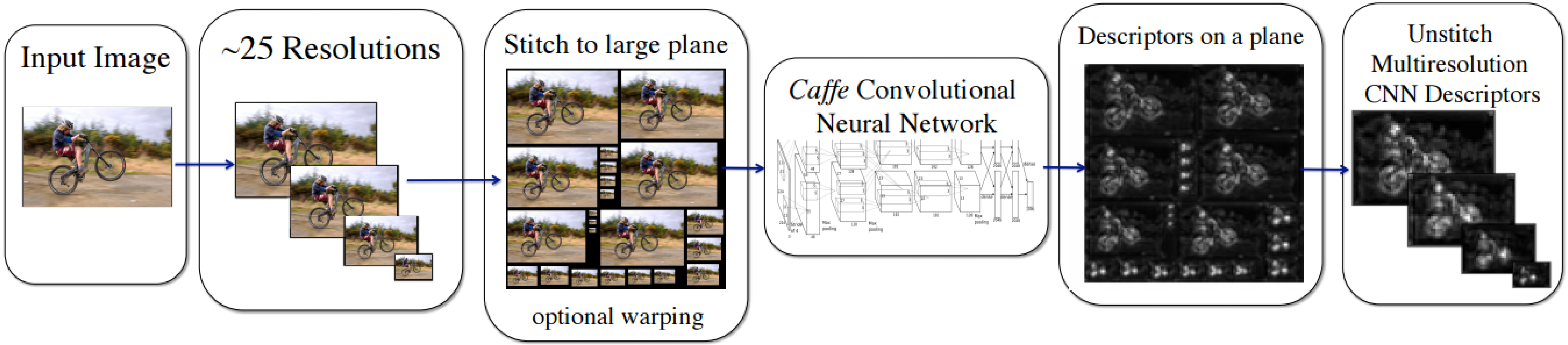


Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, Rich Feature Hierachies for Accurate Object Detection and Semantic Segmentation

Caffe: Public Domain Software



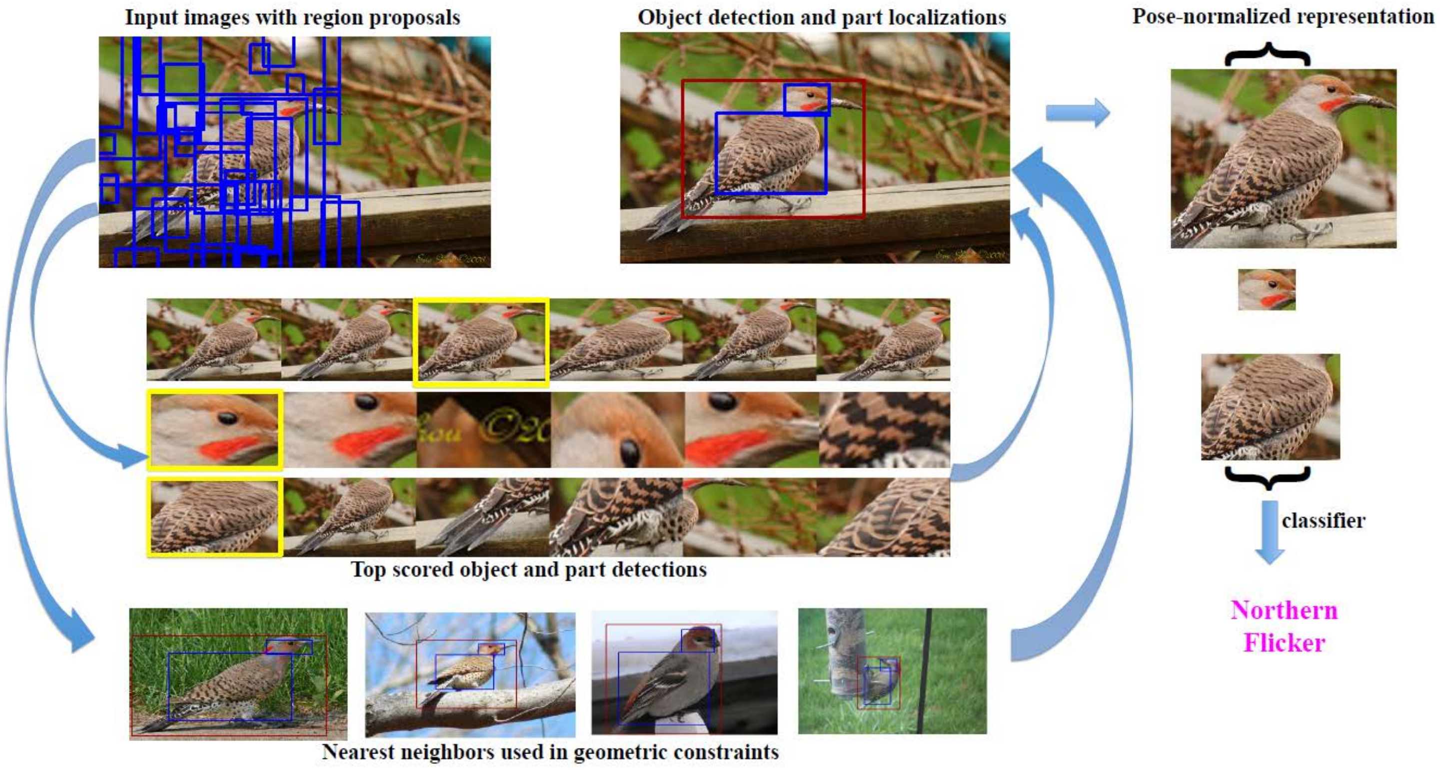
Dense Net

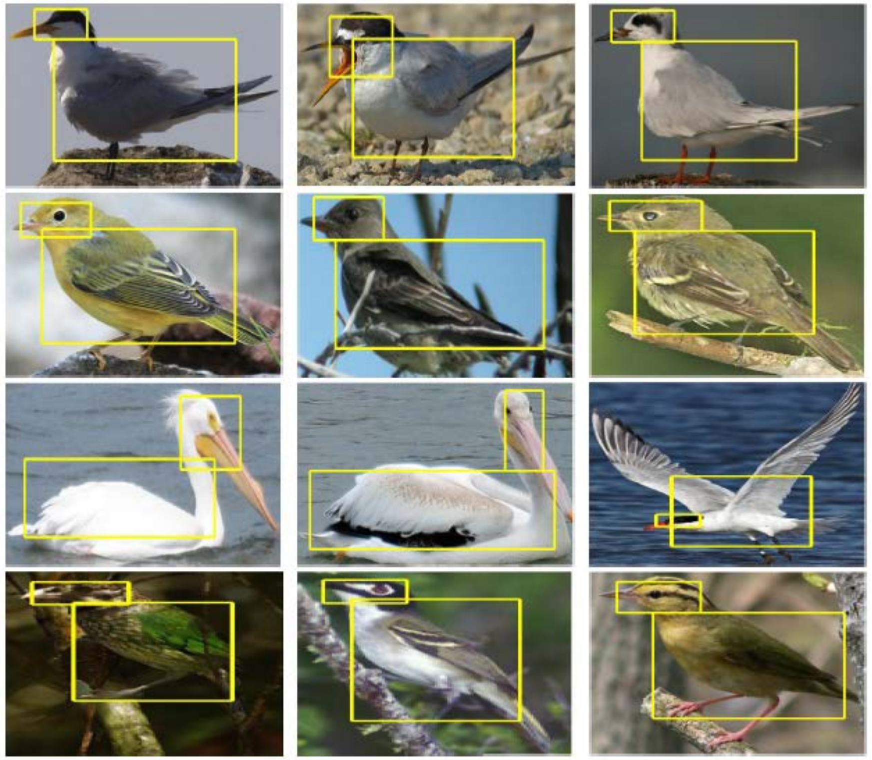


F.Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, K. Keutzer, DenseNet: Implementing Efficient ConvNet Descriptor Pyramids

Part-based CNN

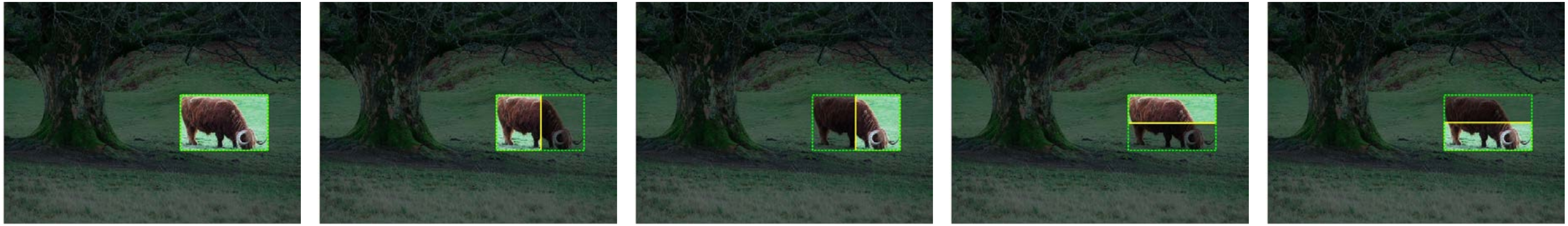
- Fine-grained Detection





Multi-region CNN

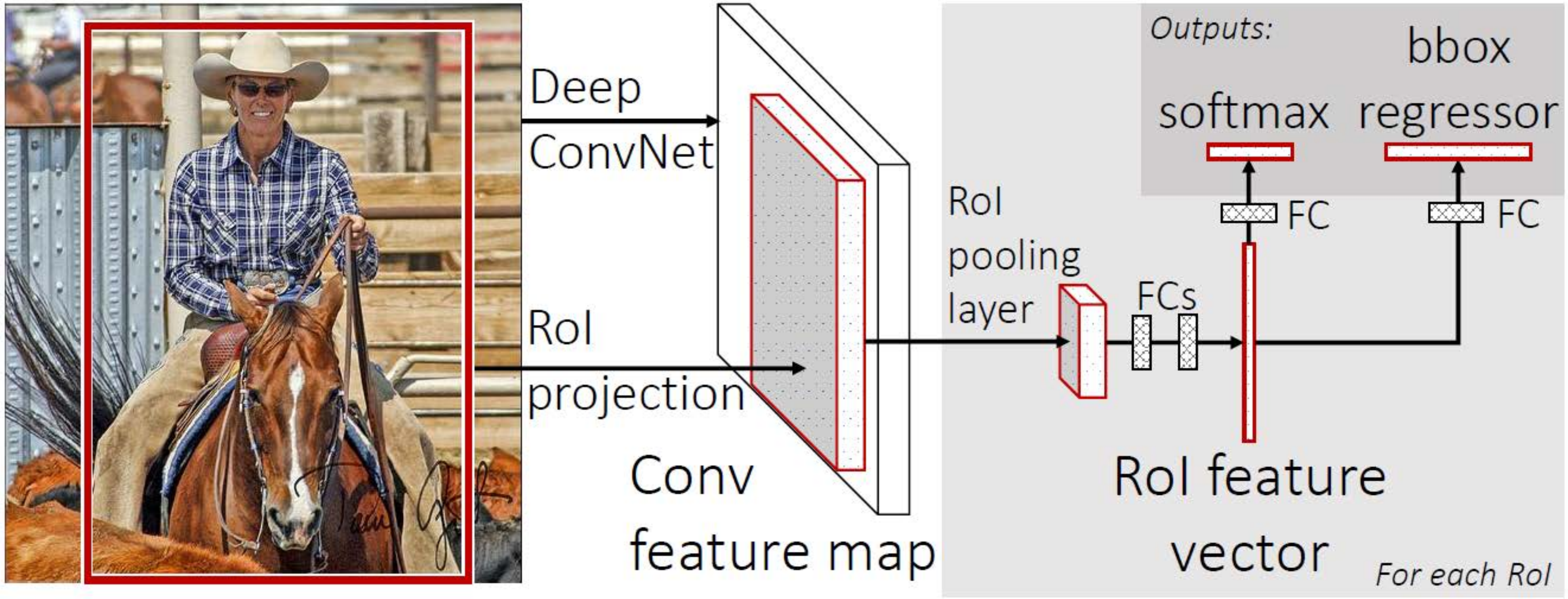
- Multi-region & semantic segmentation-aware CNN model:



(a) Original box **(b)** Half left **(c)** Half right **(d)** Half up **(e)** Half bottom

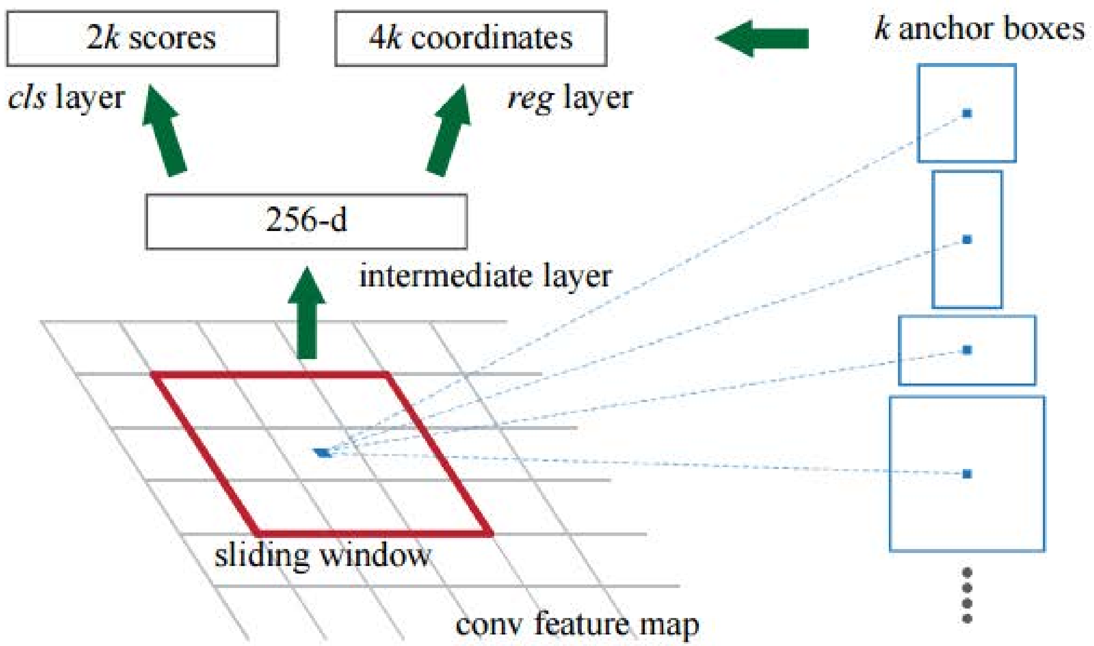
(f) Central Region **(g)** Central Region **(h)** Border Region **(i)** Border Region **(j)** Context. Region

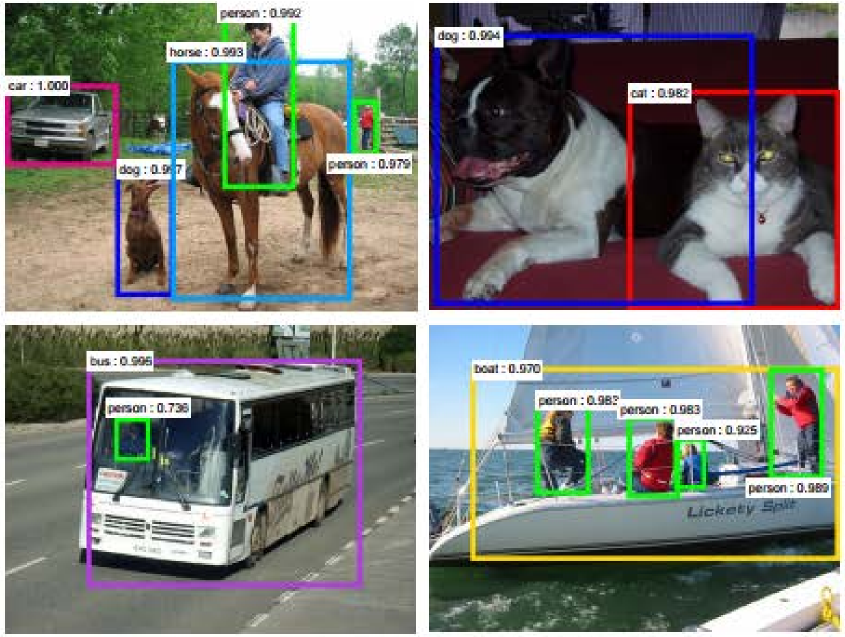
Fast RCNN



Deep
ConvNet
RoI
projection
Conv
feature map
Outputs:
bbox
softmax
regressor
FC
FC
RoI
pooling
layer
FCs
RoI feature
vector
For each RoI

Faster RCNN

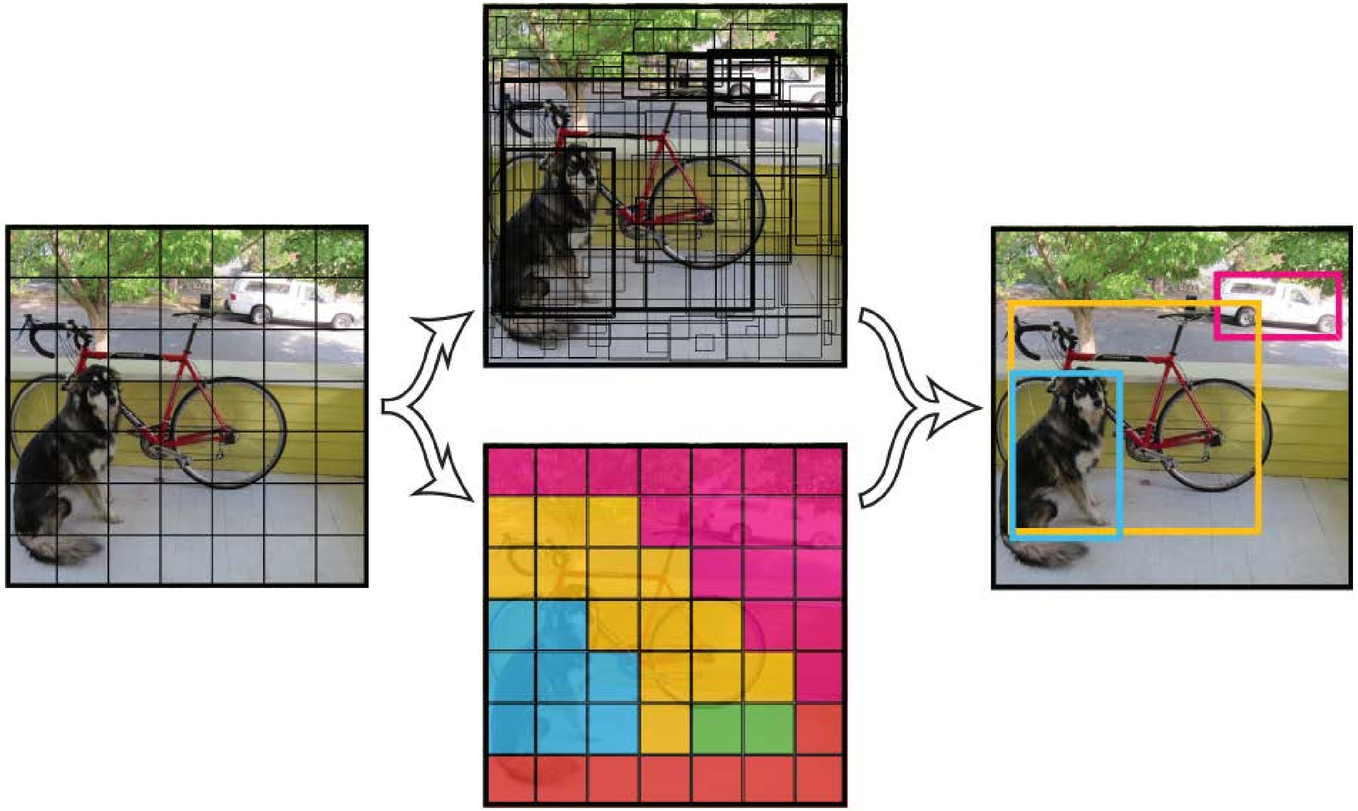




S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks.

YOLO

- More improved CNN based Object Detector:



Performance Comparison

- State-of-the-art methods comparison:

Method	ROI needed?	Regression?	Accuracy	Speed
RCNN	Y	N	62.4	0.01 fps
Fast-RCNN	Y	Y	68.4	0.01 fps
Faster-RCNN	N	Y	70.4	3 fps
YOLO	N	Y	57.9	45 fps

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

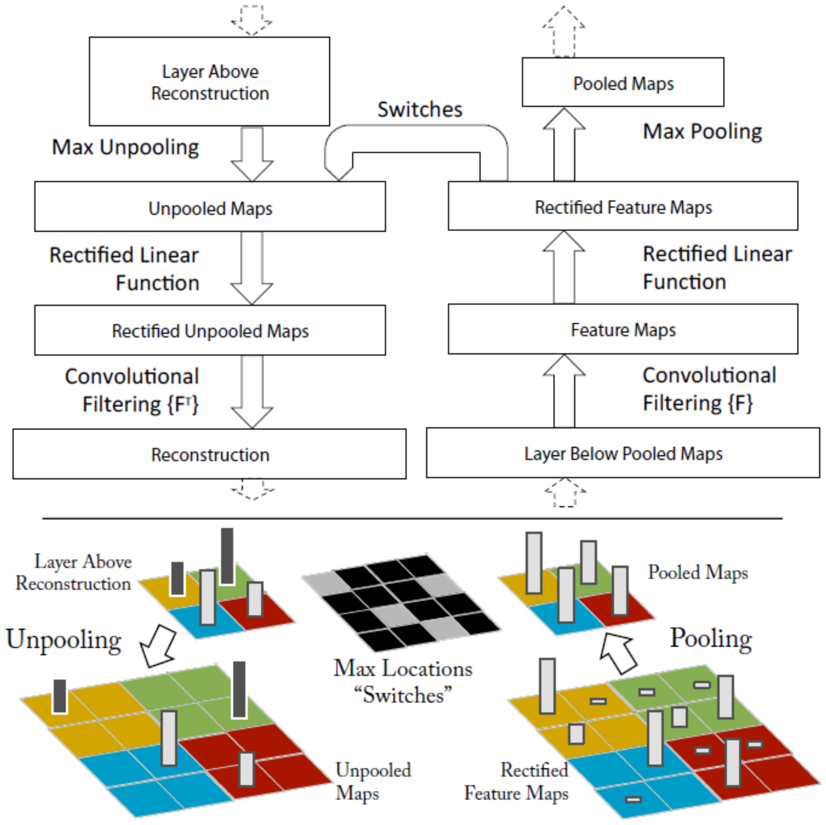
Comments

- Alex Net was the first one proposed for object detection
- Recent publications have focused on two aspects:
  - Fine-tuning for better region proposals, which serve as a critical pre-processing step for DNN
  - Seeking more applications
- Both hardware and software are accessible to the public
- However, no labeled training data are accessible to the public
- It can be implemented in the cloud platform, but not in the embedded system platform

Visualization of Deep Features

Feature Visualization

- De-conv Network:



Features at Lower Layers

- Conv1 Filter Response (Gabor-like filters)

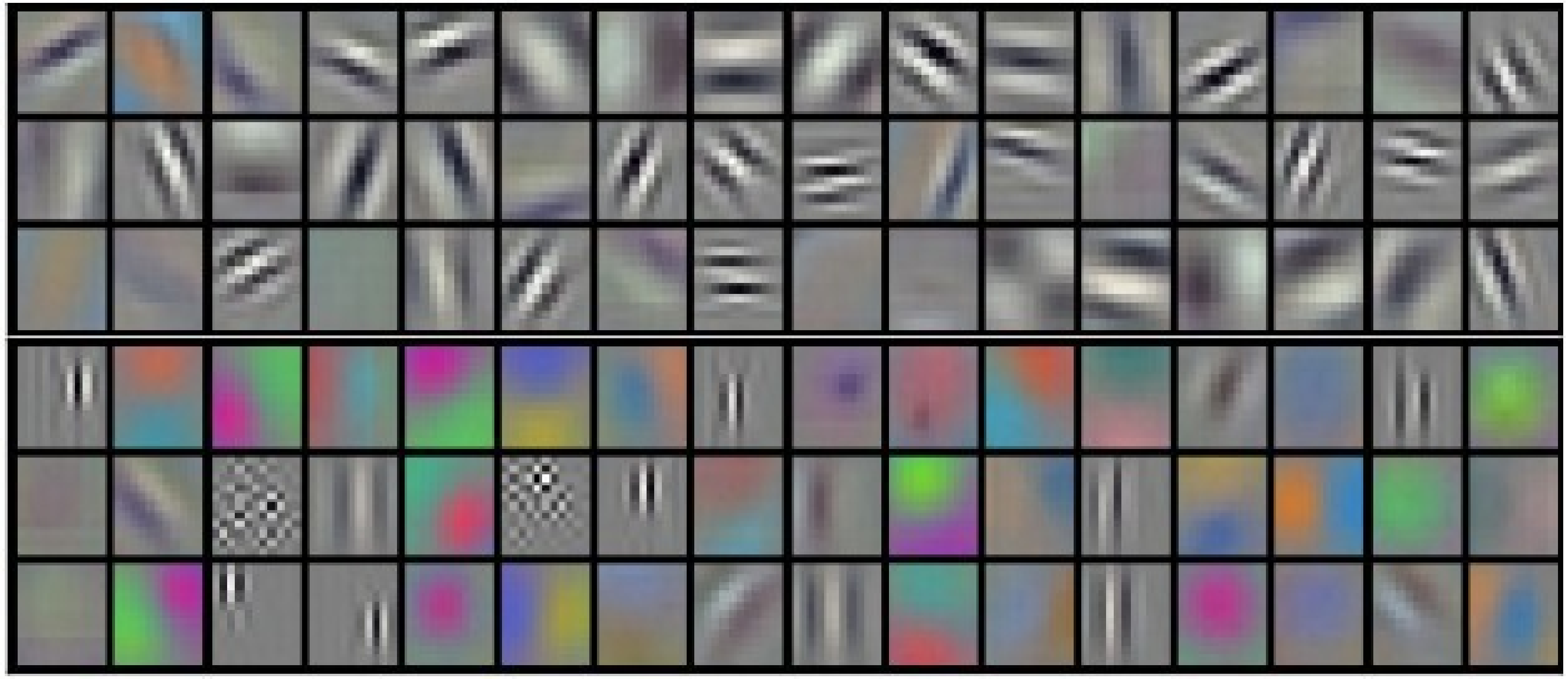
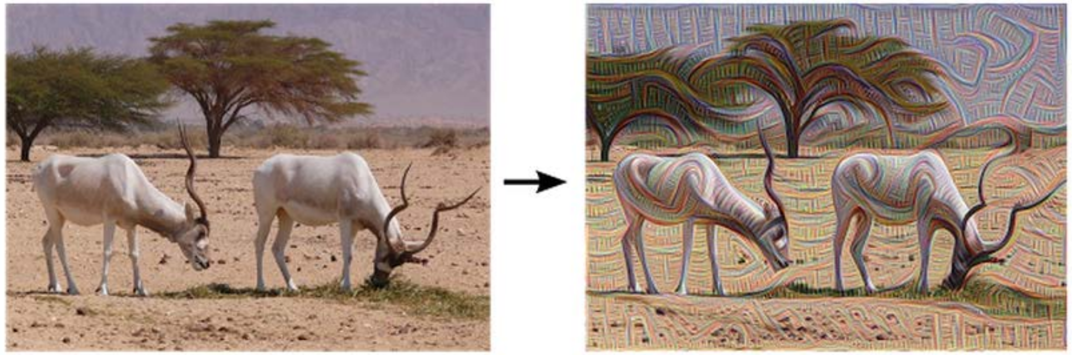
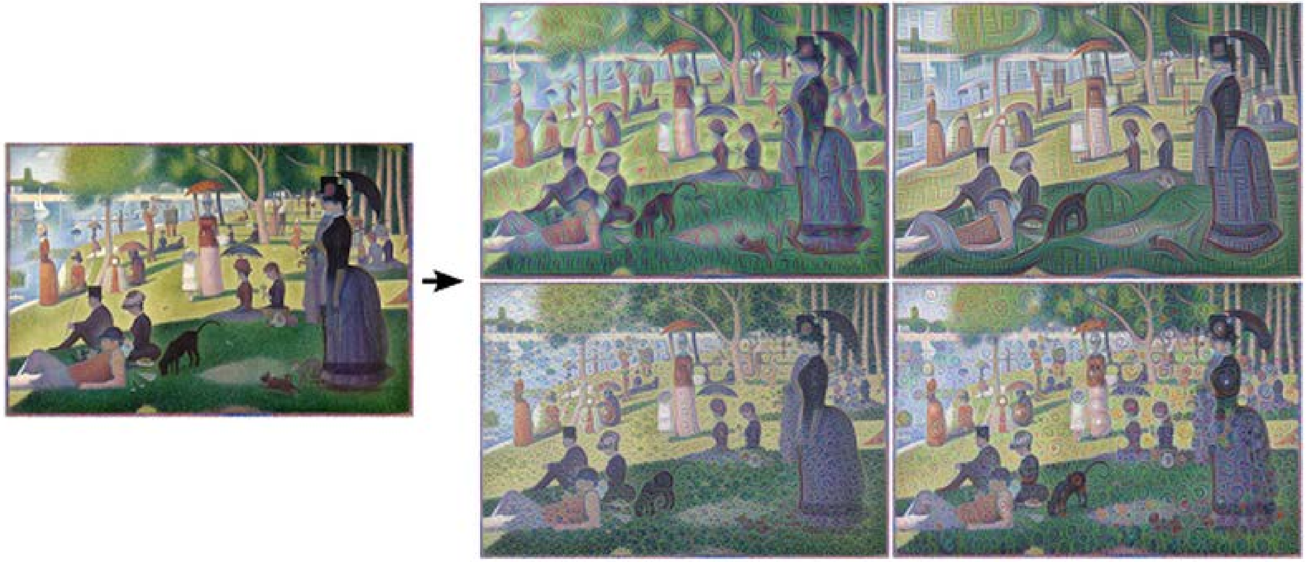


Figure Credit: Stanford CS231n Course Instruction

Reconstruction at Lower Layers

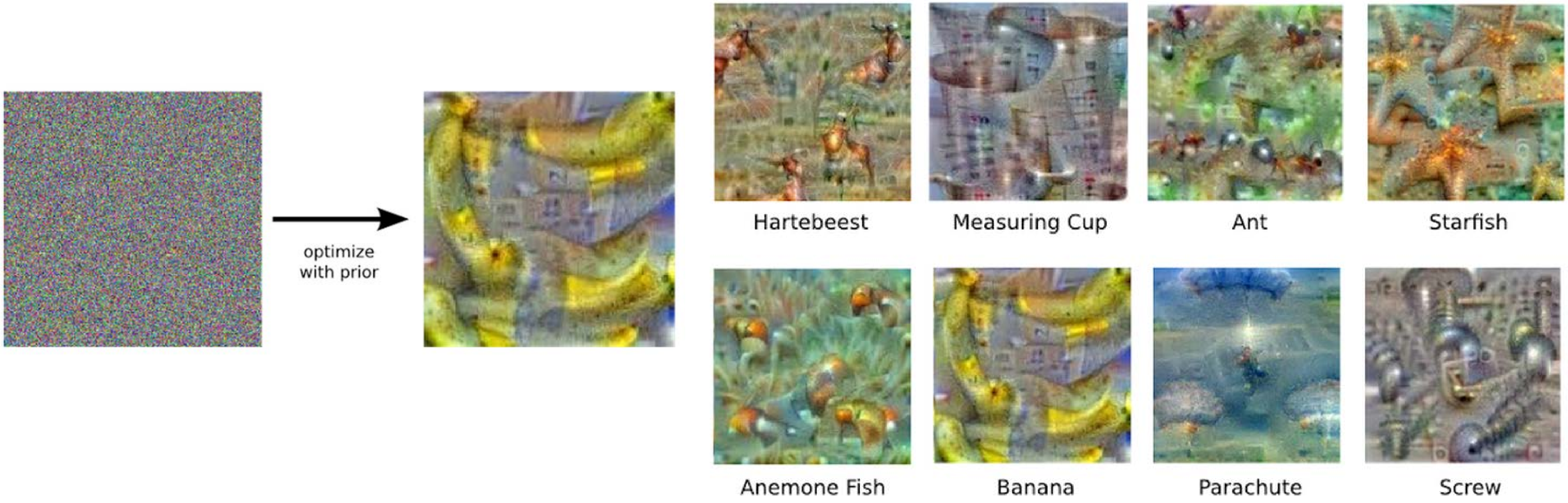
- Extracting low level features (oriented edges/contours)





Reconstruction at Higher Layers

- Random Noise Input

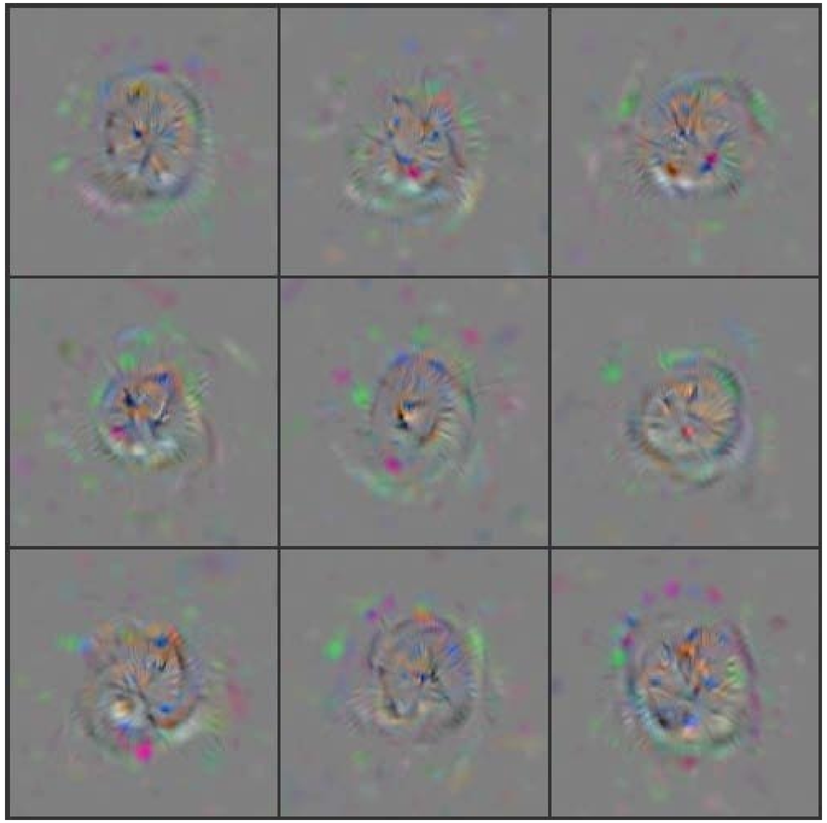


http://googleresearch.blogspot.ch/2015/06/inceptionism-going-deeper-into-neural.html

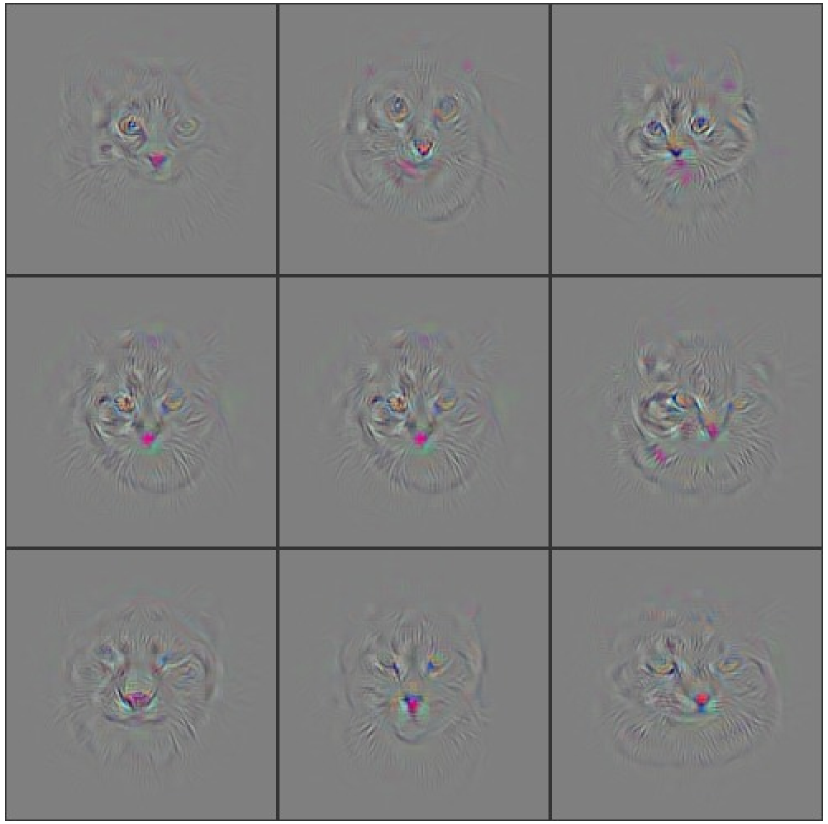
Deep Features at Conv5

- Cat Face Filter Example (Filter 111 in the Yos-CaffeNet Model):

Top 9 Input Activation Images



Max Reconstructed Input Activation



Deconv Image

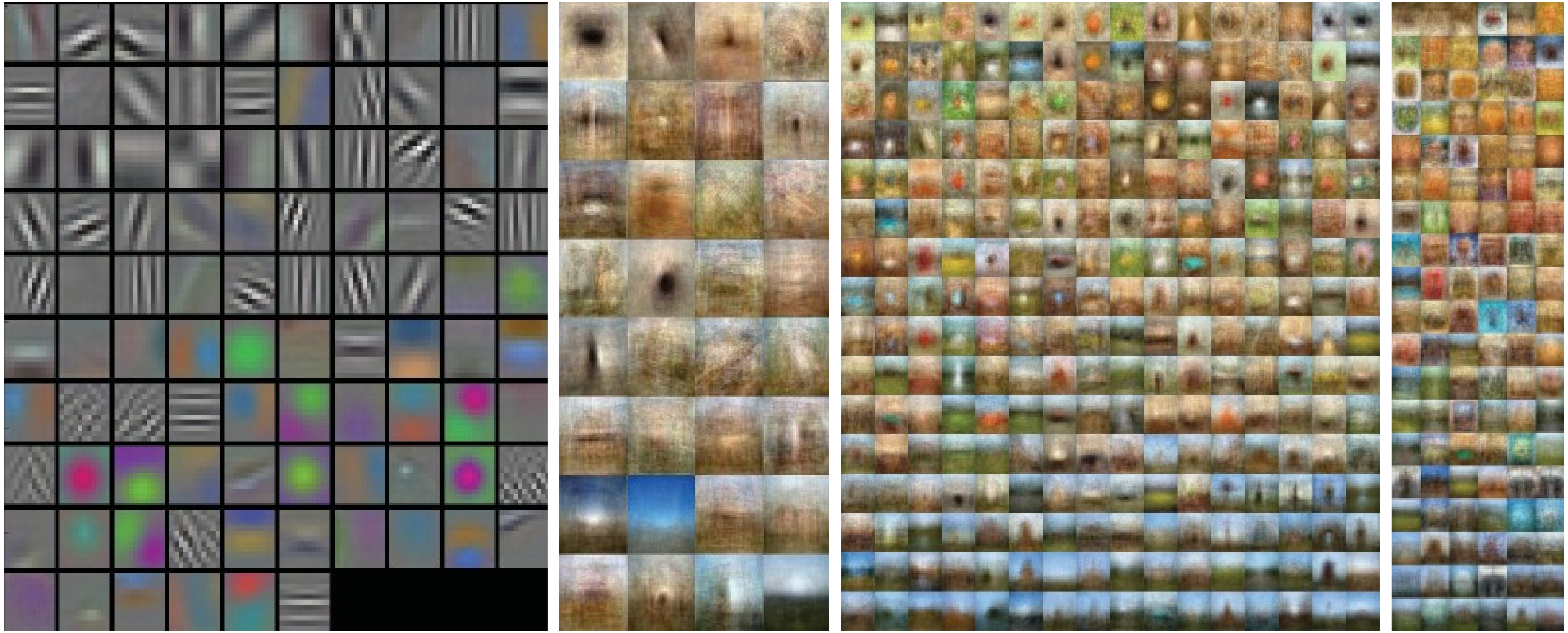
Images are generated from Deep-VIS Toolbox

Comments

- Deep feature visualization provides helpful insight into the role played by various layers
  - Lower layers – subband-filtered images with large coverage
  - Higher layers – contour-dominating images with focused coverage
  - Three general trends
    - Contour formation
      - From surfaces to contours
    - Color reduction
      - From colorful to colorless
    - Background removal
      - From larger regions (with background) to focused regions (without background)
- The whole process can be viewed as a spatial (contour) feature binding process

Deep Features for Scene Recognition

Conv 1 **Pool 2** **Pool 5** **FC 5**

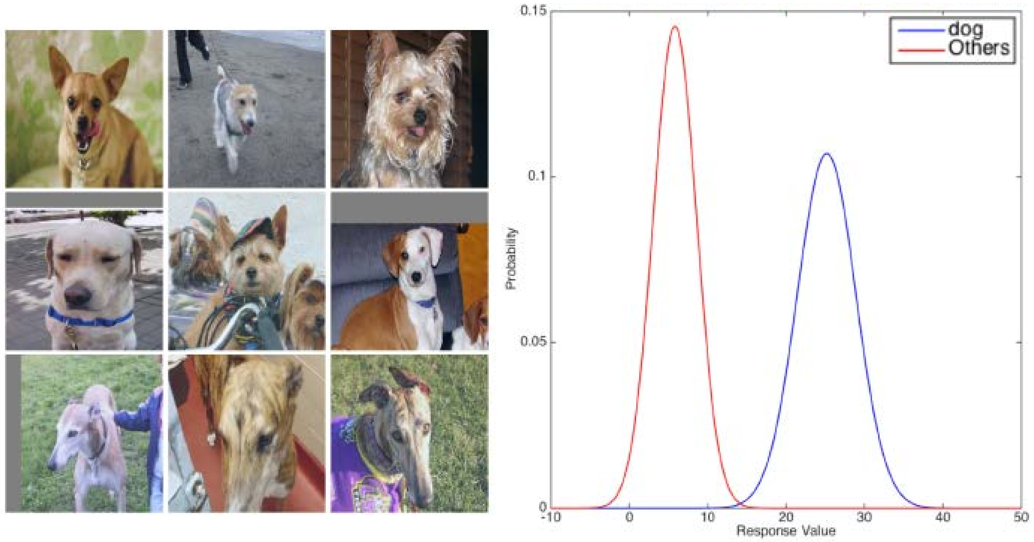


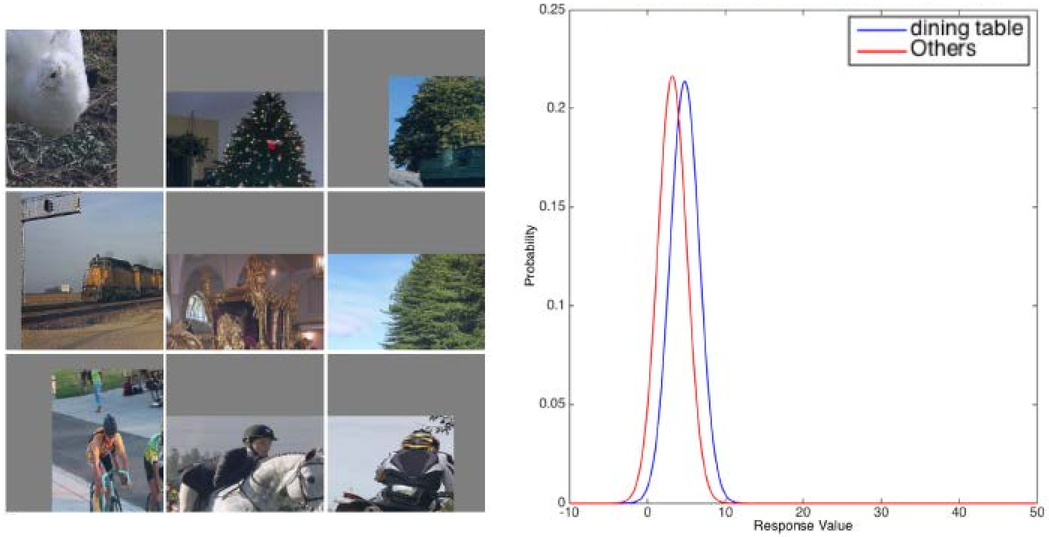
Zhou et al. “Learning deep features for scene recognition using places database”, NIPS 2014.

Gaussian Confusion Measure (GCM)

Conv5 Filter Response

- Gaussian Confusion Measure

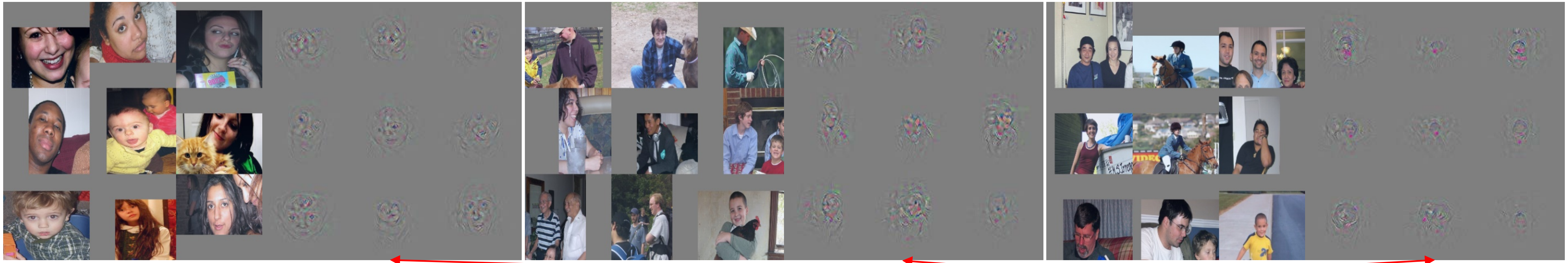


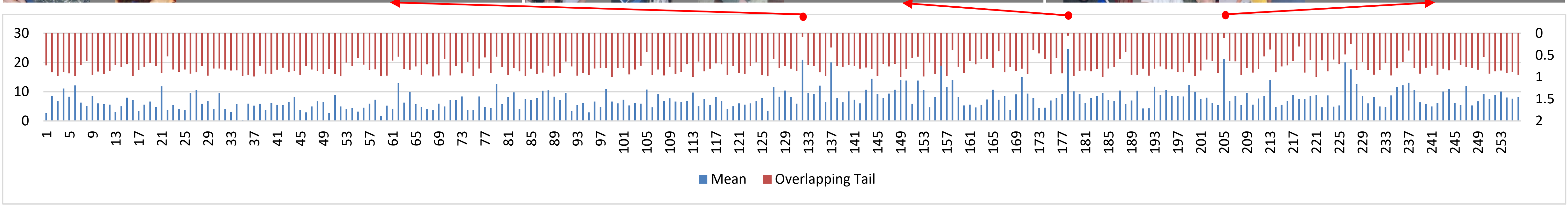


Detection Error and Mean of Conv5 Filter Responses

Responses w.r.t. human objects

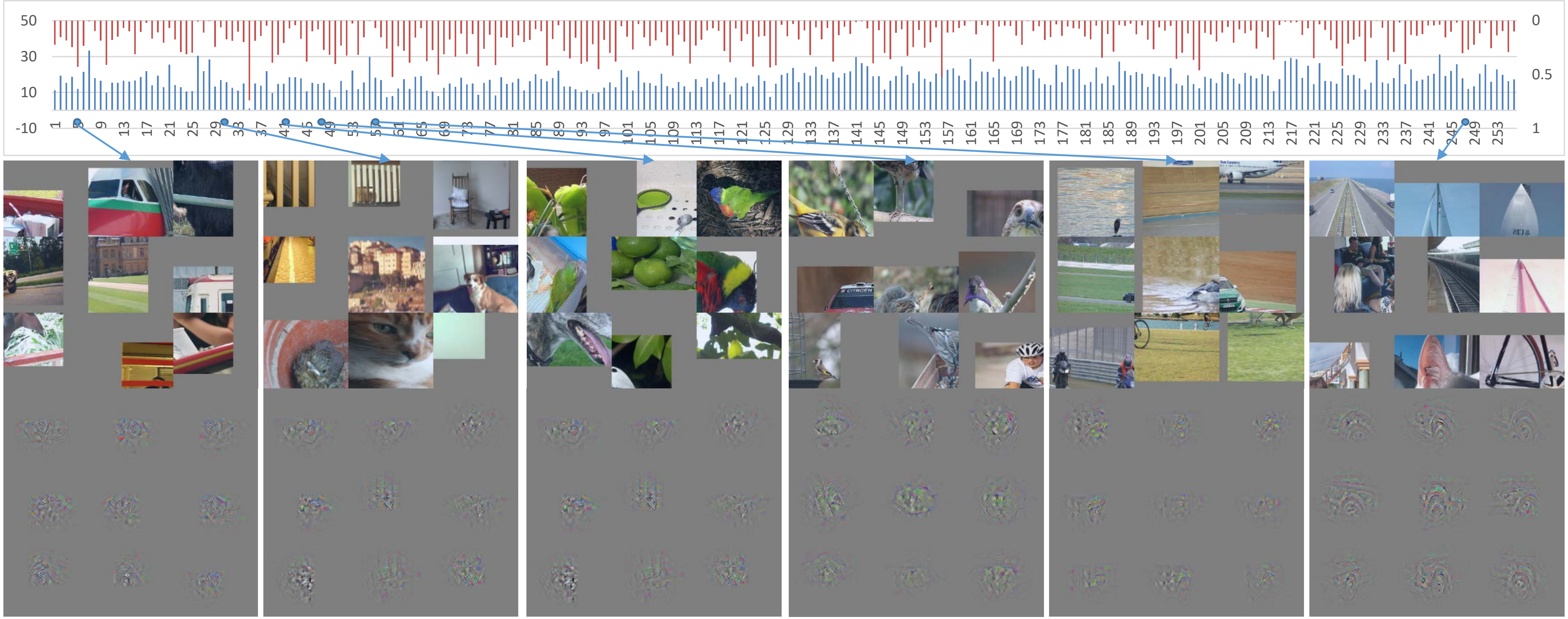
-- Known as “Grandmother Cell (GMC)” like features for human objects



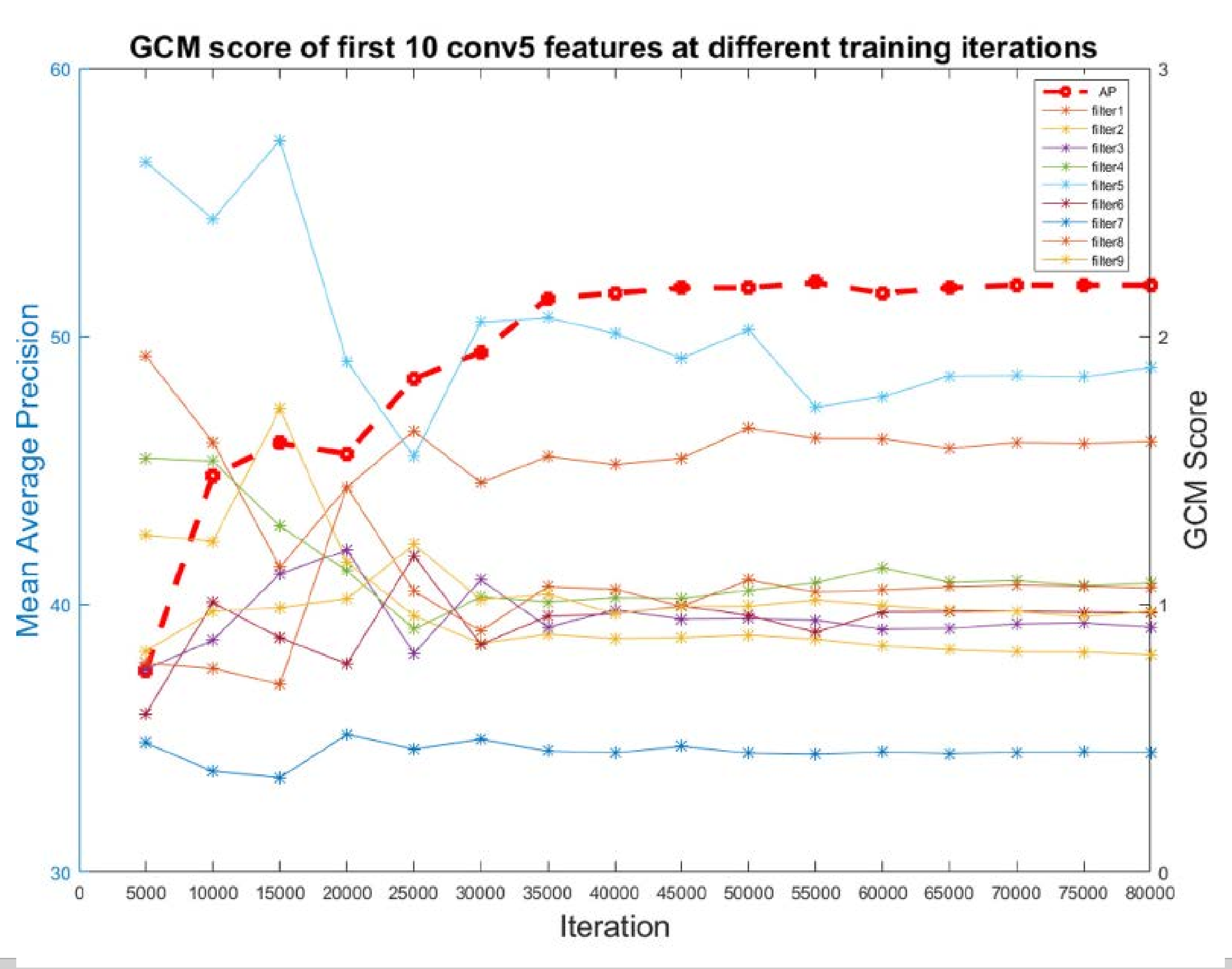


Optimal Conv5 Filters Against Multiple Object Classes (Good Examples)

Optimal Conv5 Filters Against Multiple Object Classes (Bad Examples)



GCM Scores of Top 10 Conv5 Features versus Iteration Number



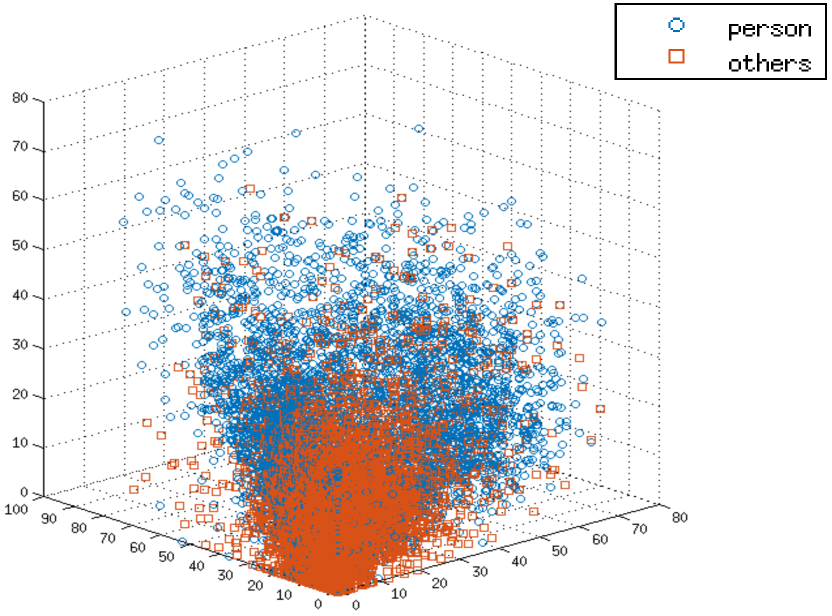
Number of GMC-like Features versus Iteration Number

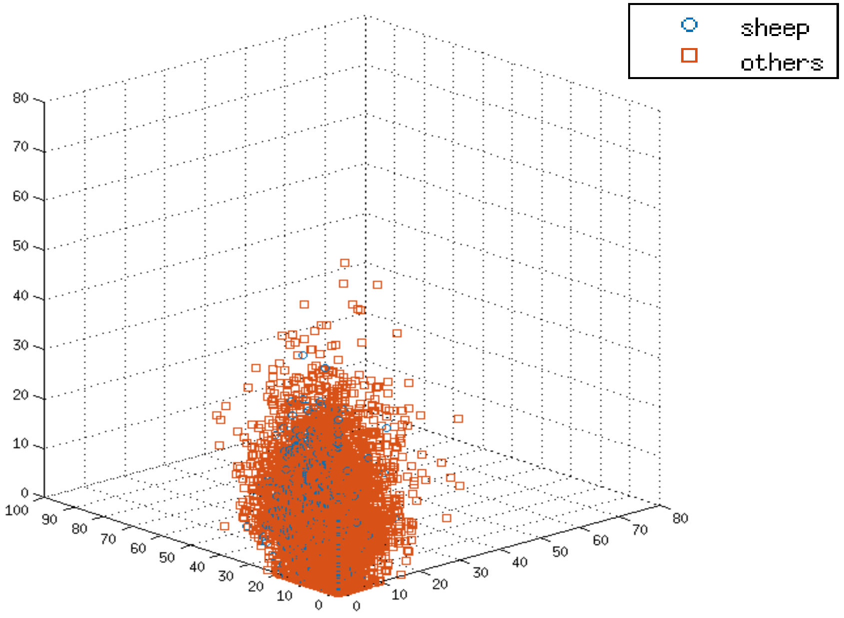
Network	Iteration	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motorbike	person	plant	sheep	sofa	train	tv	sum
Caffe	10000	4	2	0	1	0	5	0	4	0	2	1	4	4	9	3	1	1	2	1	5	49
	20000	6	3	0	0	0	5	0	8	0	2	2	9	6	10	3	1	1	2	2	5	65
	30000	6	2	0	0	0	6	0	8	0	2	3	5	4	8	3	1	1	2	1	5	57
	40000	6	2	0	1	0	8	2	7	0	2	3	6	5	9	3	2	1	3	3	5	68
	80000	5	2	0	1	0	7	2	9	0	2	5	6	5	10	3	2	1	4	4	5	73
VGG	10000	8	5	0	1	2	11	3	8	0	3	2	6	6	2	4	2	1	3	6	6	79
	20000	6	2	1	1	2	9	3	9	0	2	2	11	3	8	4	1	1	2	4	5	76
	30000	6	3	1	1	1	6	3	4	0	1	2	7	1	1	4	1	1	1	1	4	49
	40000	7	3	1	1	2	9	3	5	0	2	2	6	4	3	3	1	1	1	2	4	60
	80000	8	3	1	1	2	10	3	7	0	3	2	7	5	4	4	2	1	1	2	4	70

Cluster Purity Measure (CPM)

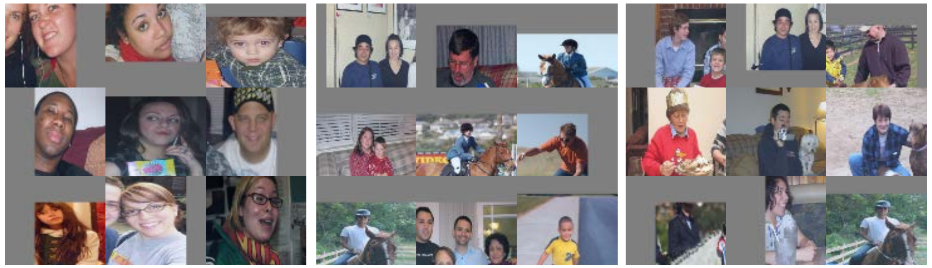
Examples of Cluster Purity Measure (1)

Good



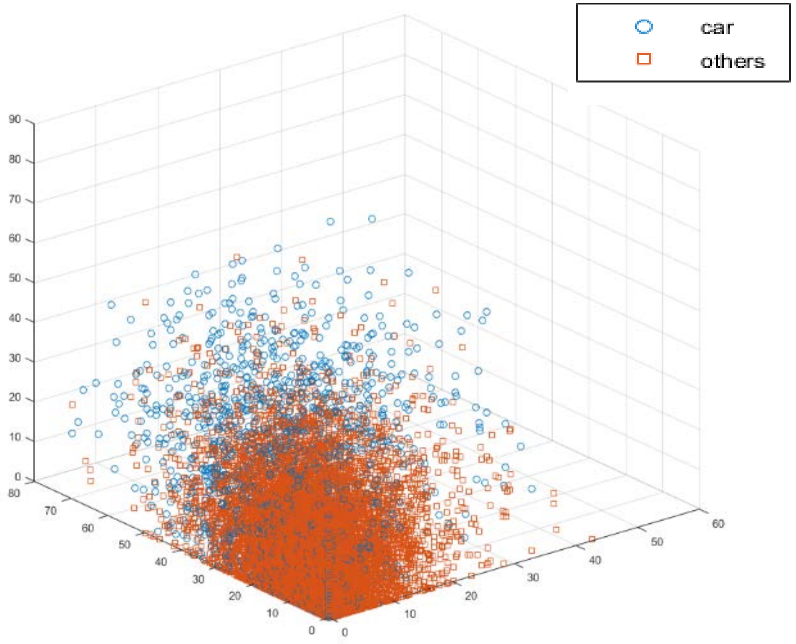


Poor

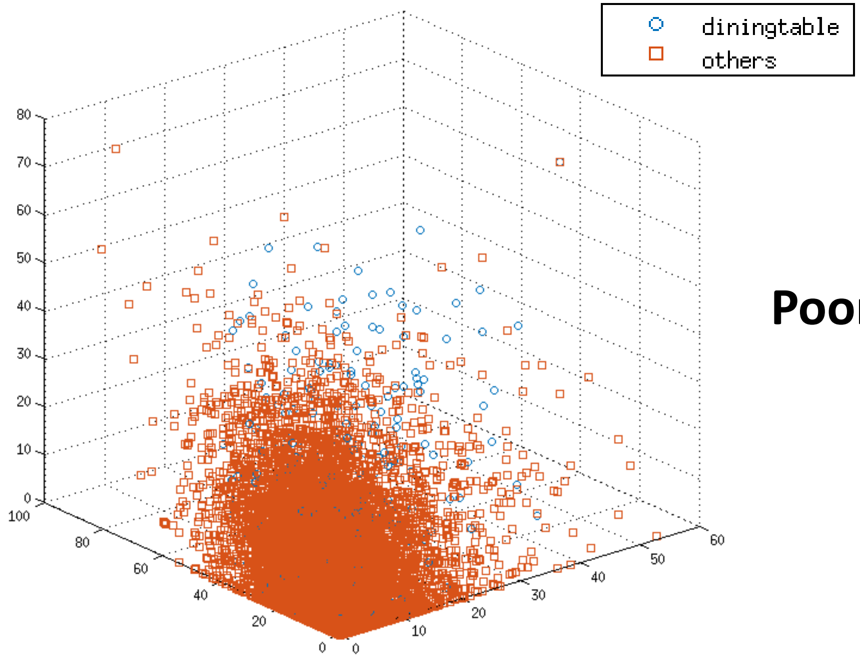


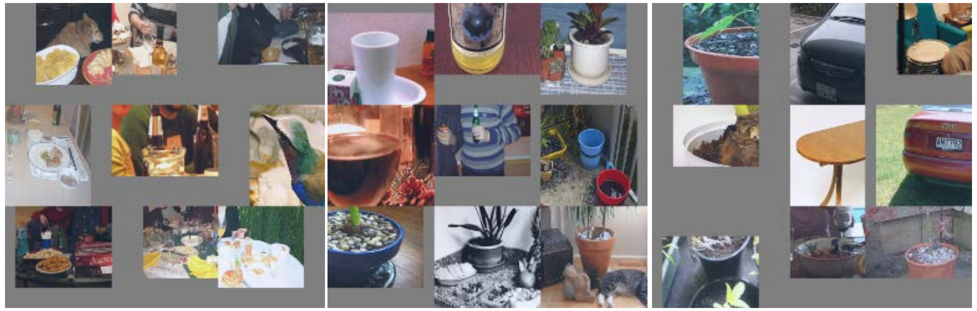
Examples of Cluster Purity Measure (2)

Good



Poor





Comments

- Besides feature visualization, we can quantify the power of a particular filter in discriminating an object class
  - 1-D case case: A smaller “Gaussian Confusion Measure (GCM)”
  - Multiple-Dimensional case: A higher “Cluster Purity Measure (CPM)”
- GCM works for a class of objects with a similar setting
- CPM works for a class of objects with multiple settings

Discussion

Why and When DNN Works Well?

- WHY: Fundamentally, a spatial domain clustering mechanism
  - Cluster images share similar spatial properties
  - Ignored in traditional pattern recognition due to limited power of image segmentation
- WHEN-1: Possible spatial combinations are limited
  - Favored object views are finite
    - Front faces
    - Side view of animals
- WHEN-2: Existence of a large number of training data
- WHEN-3: Strong correlation between training and testing datasets
  - DNN can enforce “spatial binding” of testing data by providing training data of similar nature
    - Face recognition in the wild – many companies can reach 99%

Limitations of DNN

- Fundamentally a **2D spatial binding** technique
  - The world is 3D. People can infer the depth from 2D images. DNN??
  - The world is 3D+T. It is still difficult for DNN to handle video
  - Needs the support of object proposal
    - The whole process (object proposal + object recognition) appears to be a brute force (or detour) solution.
- Engineering cost
  - The cost of training data collection
  - The computing cost
  - These two are not fundamental limitations

My Perspectives

- DNN conducts temporal binding to achieve speech recognition
  - “1D signal nature + linguistic characteristics” imposes a bound on speech variability
  - DNN can offer a powerful solution
- DNN conducts spatial binding to achieve object recognition
  - The complexity of scene arrangement is much higher than 1-D speech arrangement
  - It may make sense for some industrial companies to adopt it to solve a niche problem. However, it is a different story for academia
- My perspectives:
  - **DNN is still a rapidly growing field. The performance is outstanding, yet more understanding is needed.**