Deep Nets - What they have ever done for vision?

Sebastian Cygert

Deep Nets: What have they ever done for Vision?*

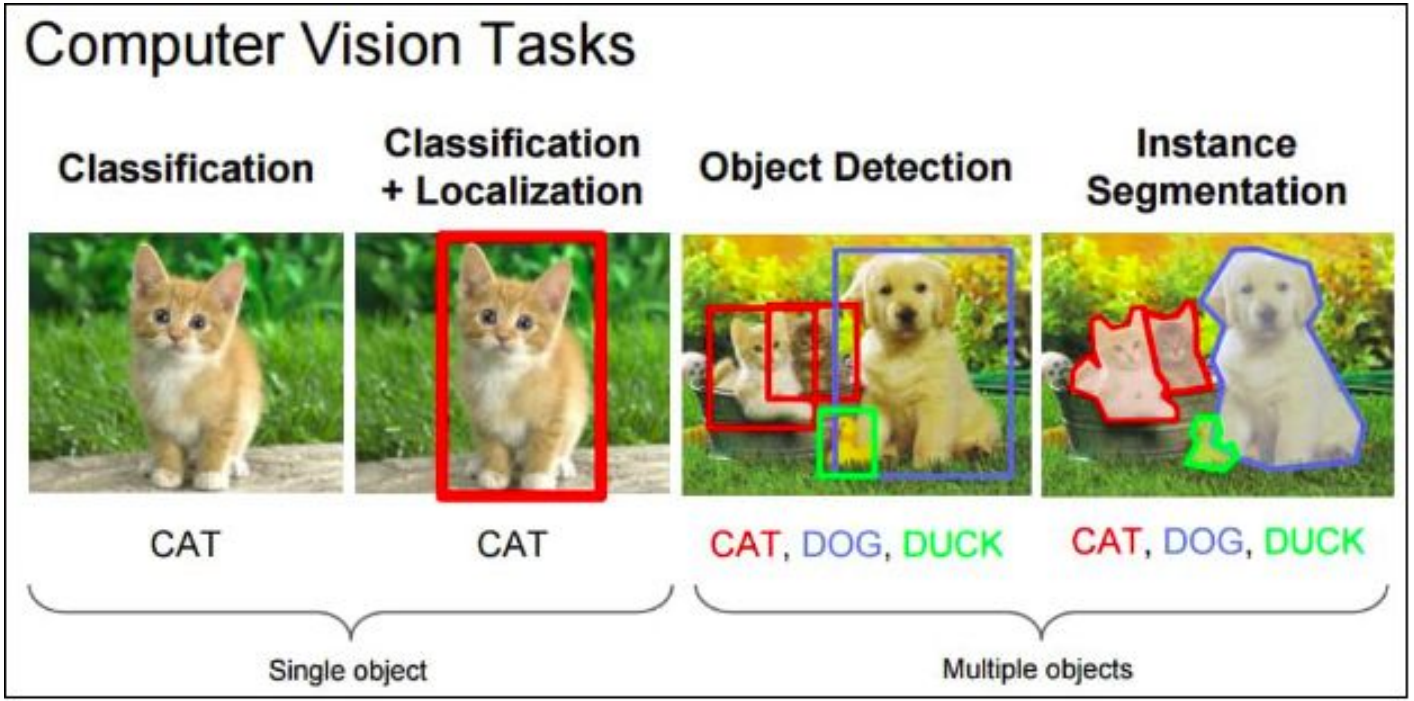
Alan L. Yuille[1,2] Chenxi Liu[2]
Department of Cognitive Science[1] & Computer Science[2]
Johns Hopkins University

May 11, 2018

This is an opinion paper about the strengths and weaknesses of Deep Nets. They are at the center of recent progress on Artificial Intelligence and are of growing importance in Cognitive Science and Neuroscience since they enable the development of computational models that can deal with a large range of visually realistic stimuli and visual tasks. They have clear limitations but they also have enormous successes. There is also gradual, though incomplete, understanding of their inner workings. It seems unlikely that Deep Nets in their current form will be the best long-term solution either for building general purpose intelligent machines or for understanding the mind/brain, but it is likely that many aspects of them will remain. At present Deep Nets do very well on specific types of visual tasks and on specific benchmarked datasets. But Deep Nets are much less general purpose, flexible, and adaptive than the human visual system. Moreover, methods like Deep Nets may run into fundamental difficulties when faced with the enormous complexity of natural images. To illustrate our main points, while keeping the references small, this paper is slightly biased towards work from our group.

Overview

1. Success story.
2. Understanding deep nets.
3. Training with less supervision / transfer learning.
4. What does not work yet.
5. Compositional models.

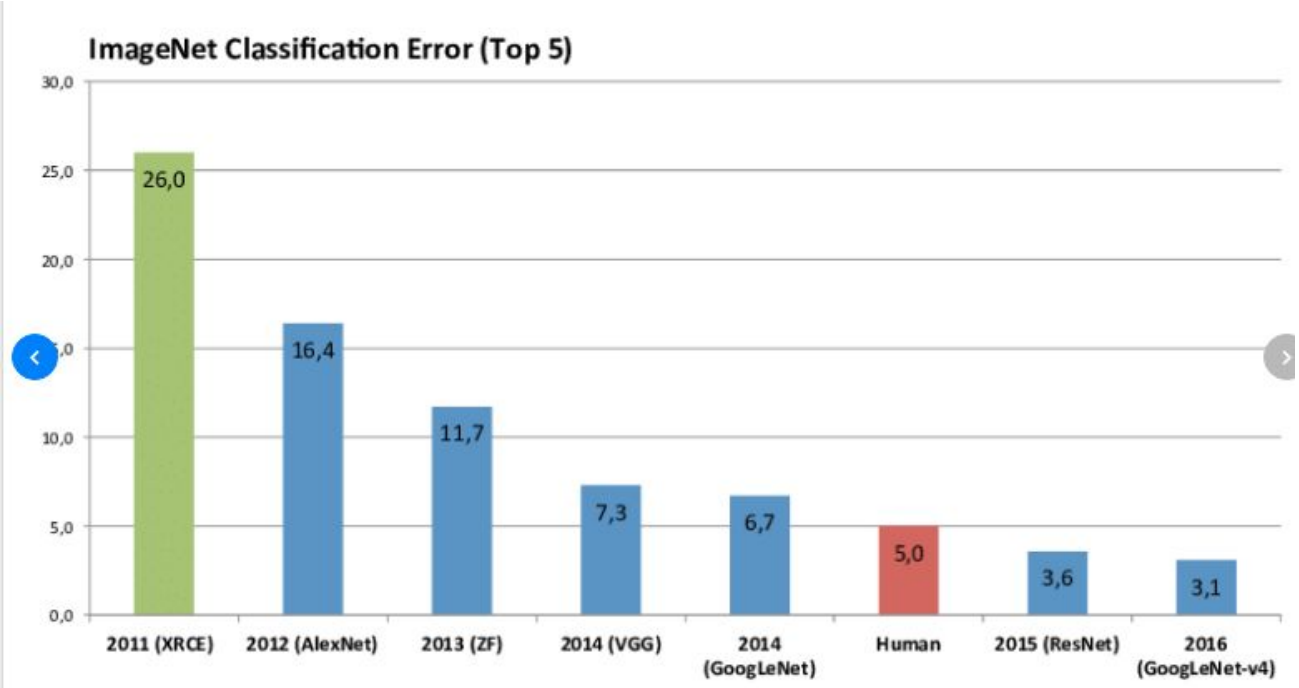


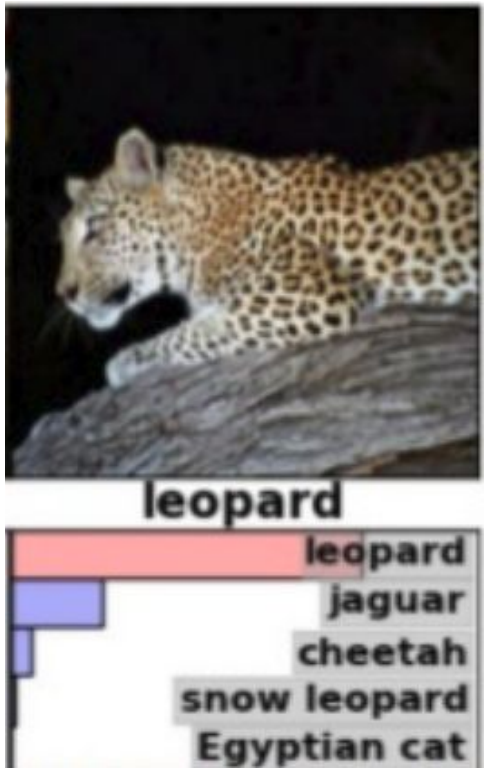
Supervised training - huge number of specialized datasets: ImageNet, MSCOCO (general categories), KITTI (autonomous cars), UA-Detrac (traffic monitoring), ...

http://www.themtank.org/a-year-in-computer-vision

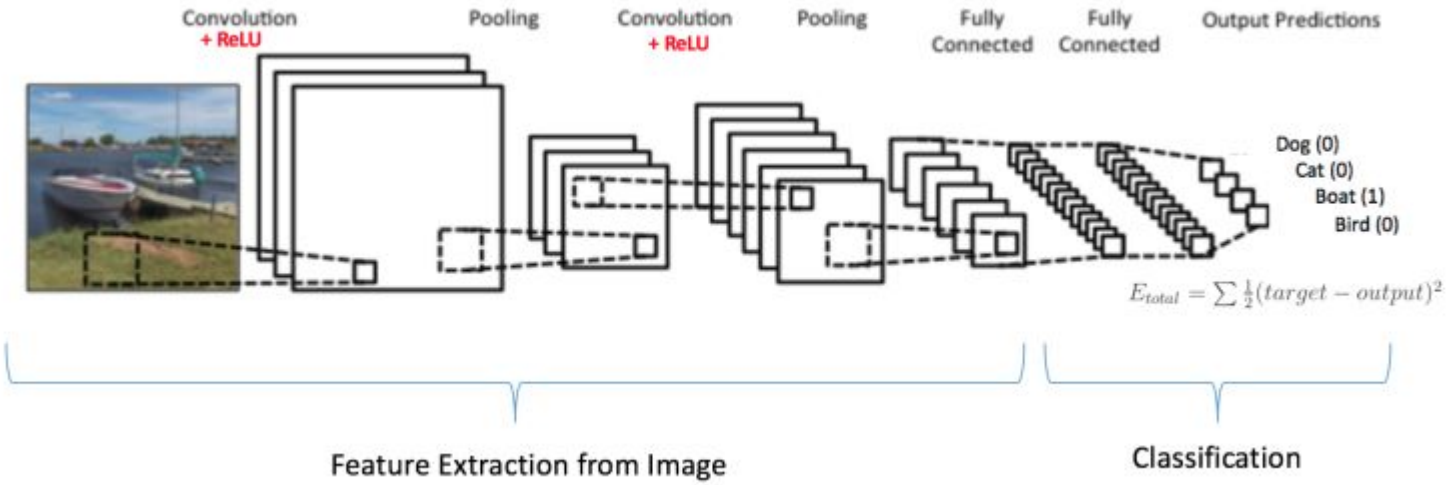
ImageNet

Categorize image to one of 1000 categories. 1.2M images in training set, 100K in test. Deep learning revolution starting in 2012



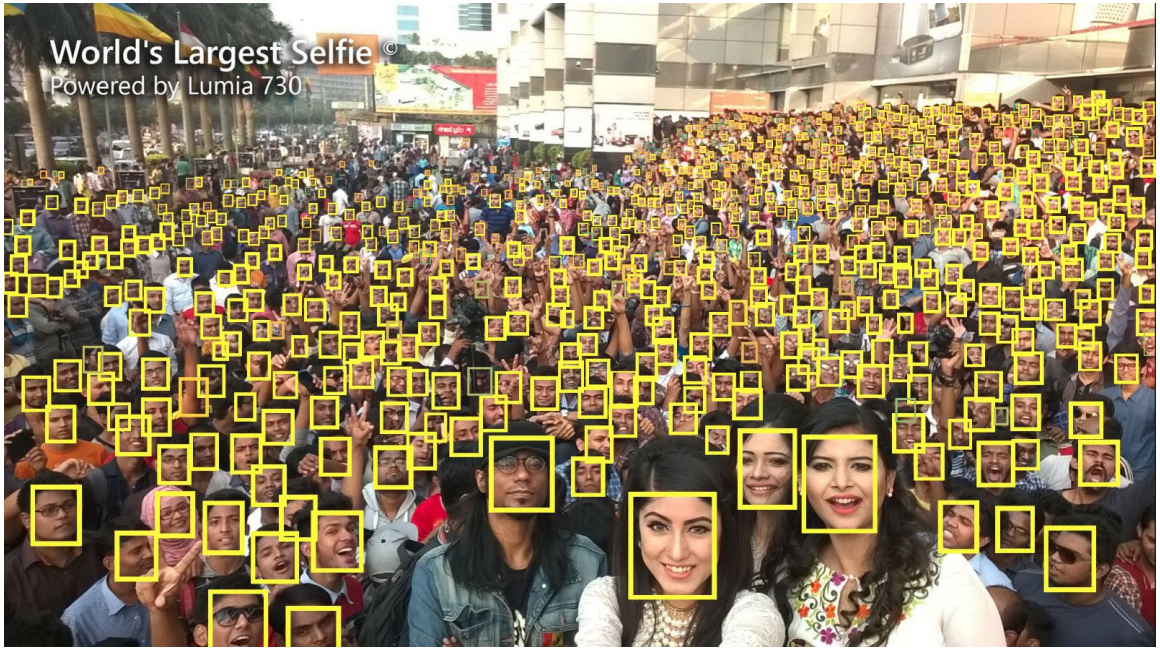


DL architecture for CV



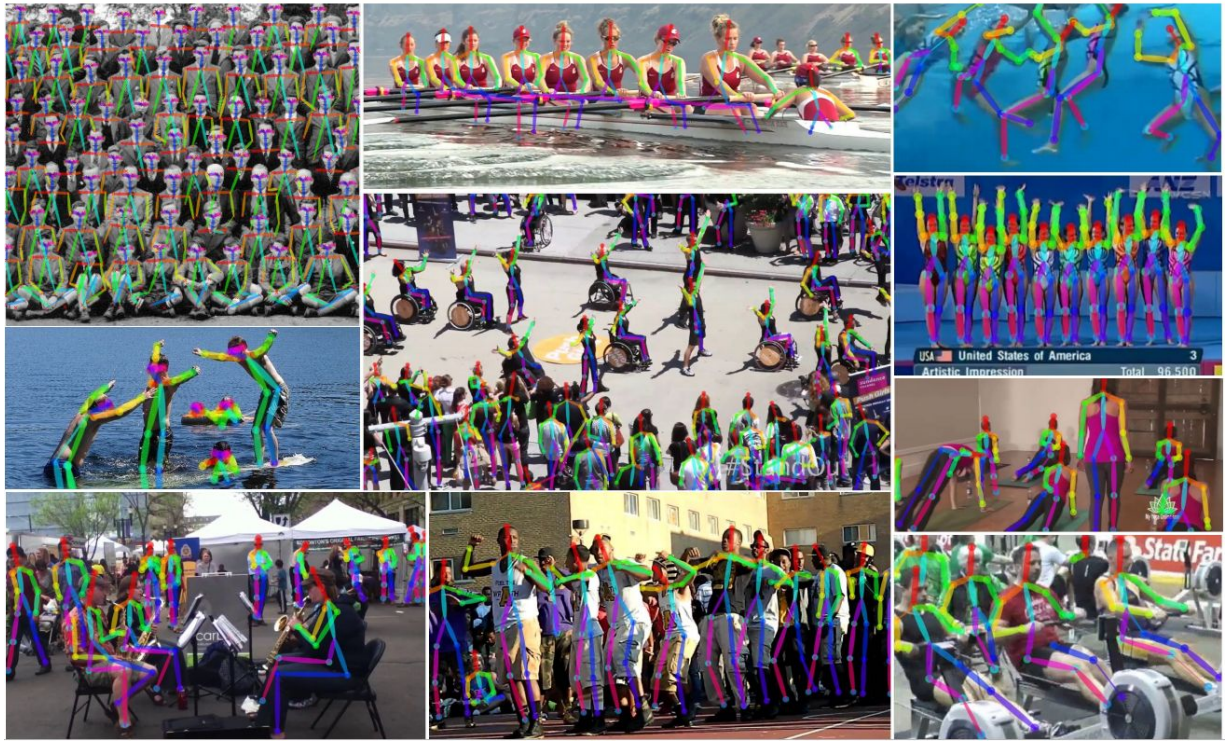
- Learning powerful (semi-hierarchical) image representation (instead of hand crafted ones)
- Scalability to millions of training examples

Impressive performance on many tasks



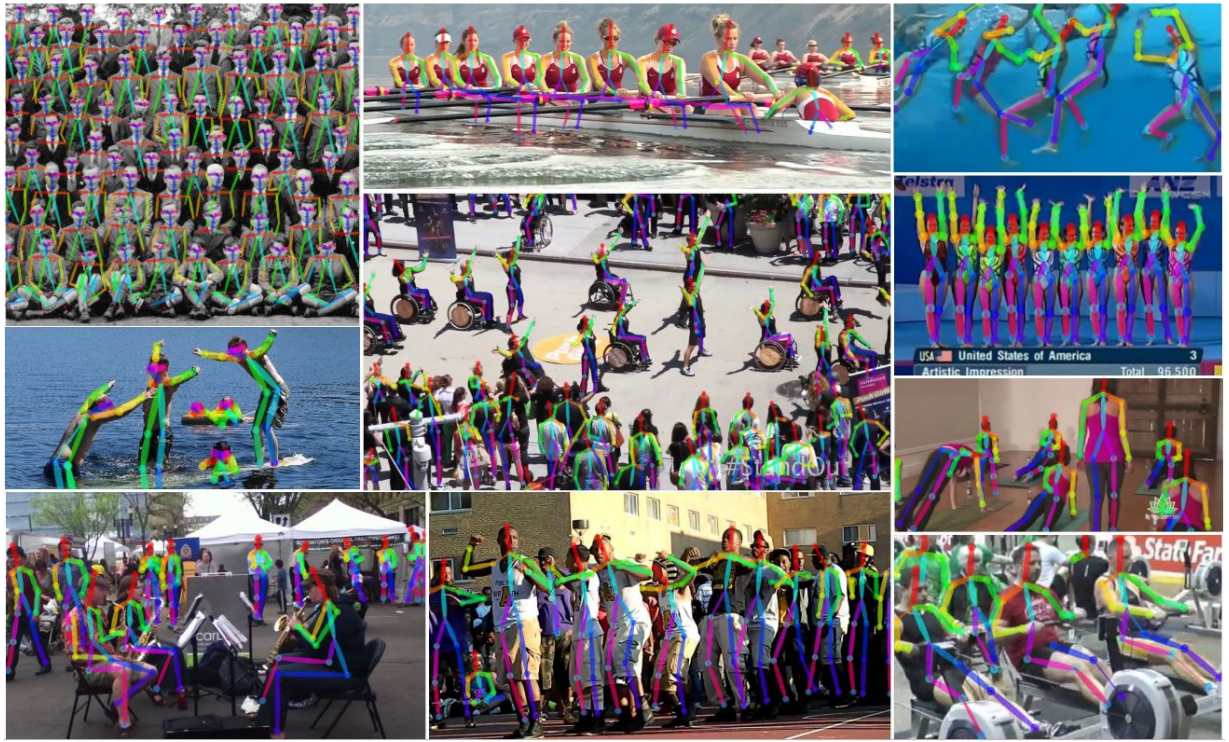
P. Hu, D. Ramanan, Finding Tiny Faces, 2016, CMU

Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields



MPII Human Pose dataset of articulated human pose estimation. The dataset includes around **25K images** containing over **40K people** with annotated body joints.

Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields

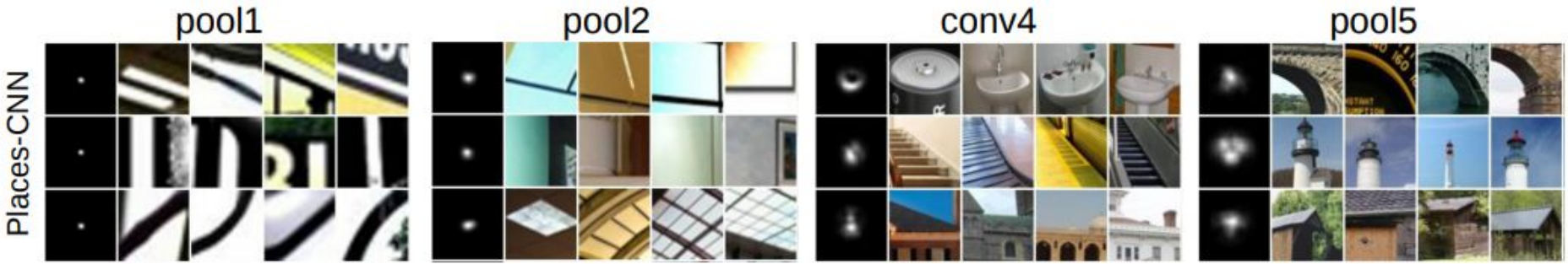


MPII Human Pose dataset of articulated human pose estimation. The dataset includes around **25K images** containing over **40K people** with annotated body joints.

Understanding deep nets

Understanding Deep Nets

Neural network trained to classifier scenes
High level layer appears to be detecting semantic objects



Object detectors emerge in deep scene CNNs, A. Torralba et al. 2015

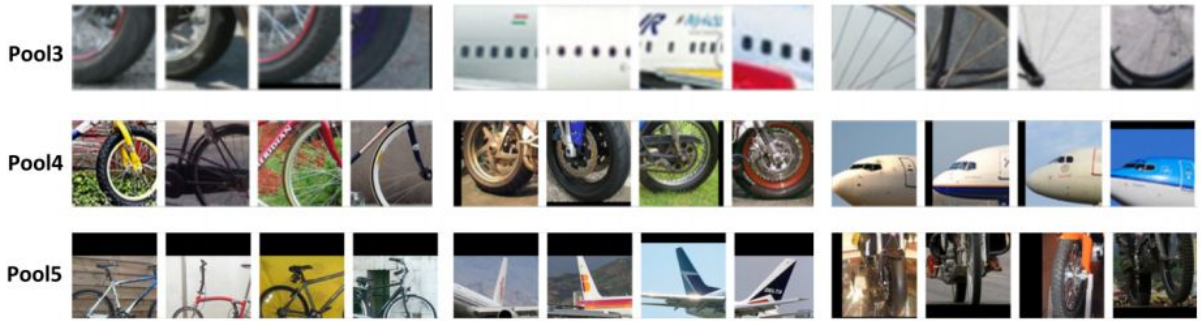


Figure 4: Figure taken from Wang et al. (2015). The visual concepts obtained by population encoding are visually tight and we can identify the parent object class pretty easily by just looking at the mid-level concepts.

Unsupervised learning of object semantic parts from internal states of CNNs by population encoding

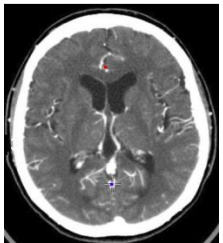
Transfer learning

- Disadvantage of Deep Nets is that they typically need a very large amount of annotated (i.e. fully supervised) training data

Transfer learning:

- pre-train neural network weights on similar dataset (ImageNet?) to obtain good image / feature representation
- fine-tune with limited annotated set

What if we want to apply DL to problem where there is no related annotated dataset (i.e. CT scans, robotics)?

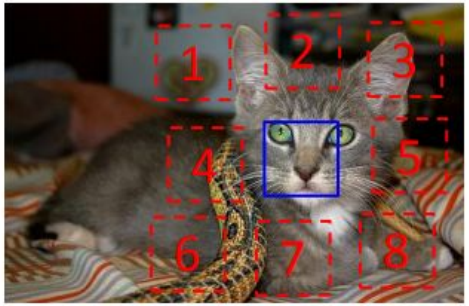


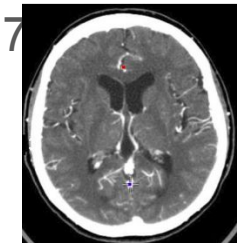
Self-supervised learning

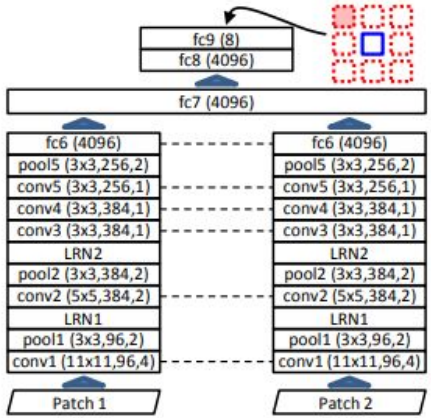
Humans learns a lot in unsupervised manner

Learn **image representation** without annotations by using pretext task on **large scale** datasets:

- Unsupervised Visual Representation Learning by Context Prediction (C. Doersch et al. 2015)
- **Self-supervised learning** + **reinforcement learning (**Curiosity-driven Exploration by Self-supervised Prediction, D. Pathak, 2017



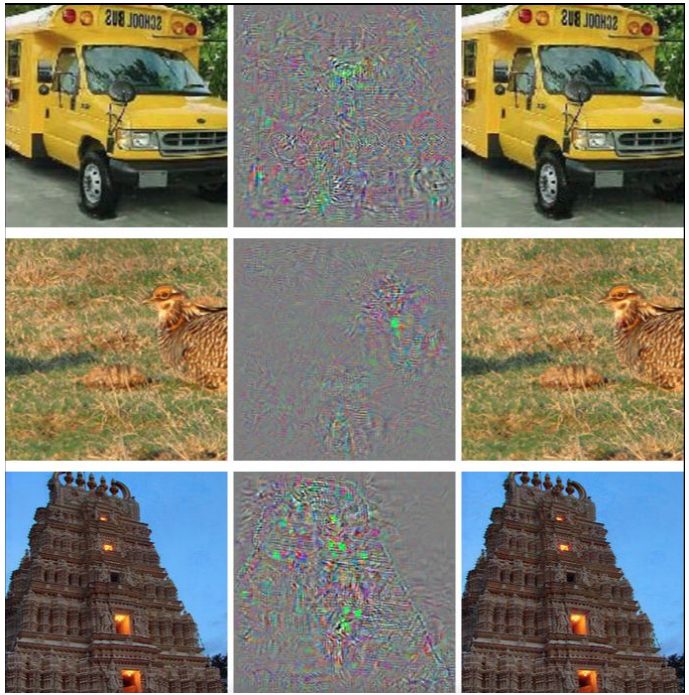


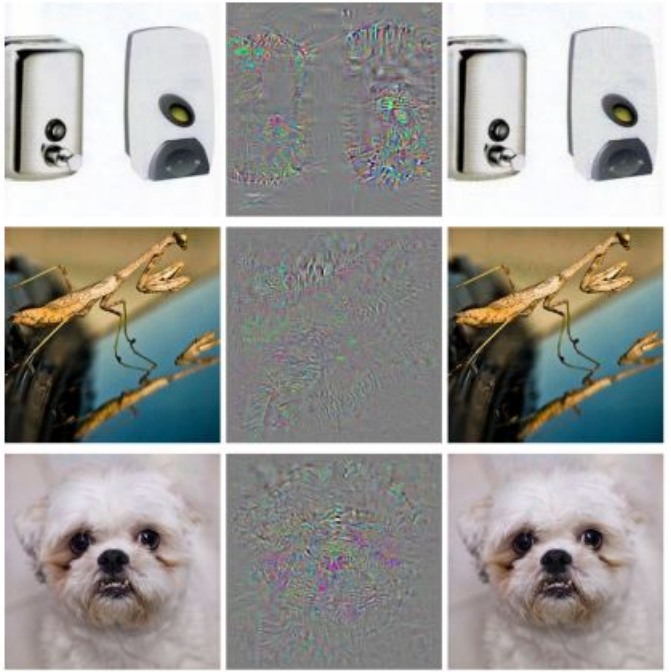


What does not work

- Adversarial examples
- Robustness
- Generalization
- Learning for the long tail
- Safety critical applications

Adversarial examples - classification at pixel level





All images on the right classified as **ostrich**

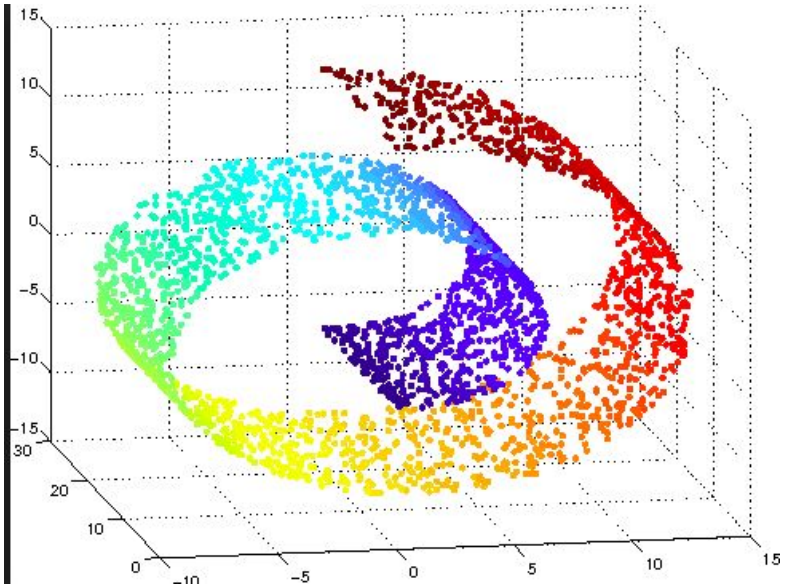
C. Szegedy et. al. Intriguing properties of neural networks, 2014

Adversarial examples - intuition

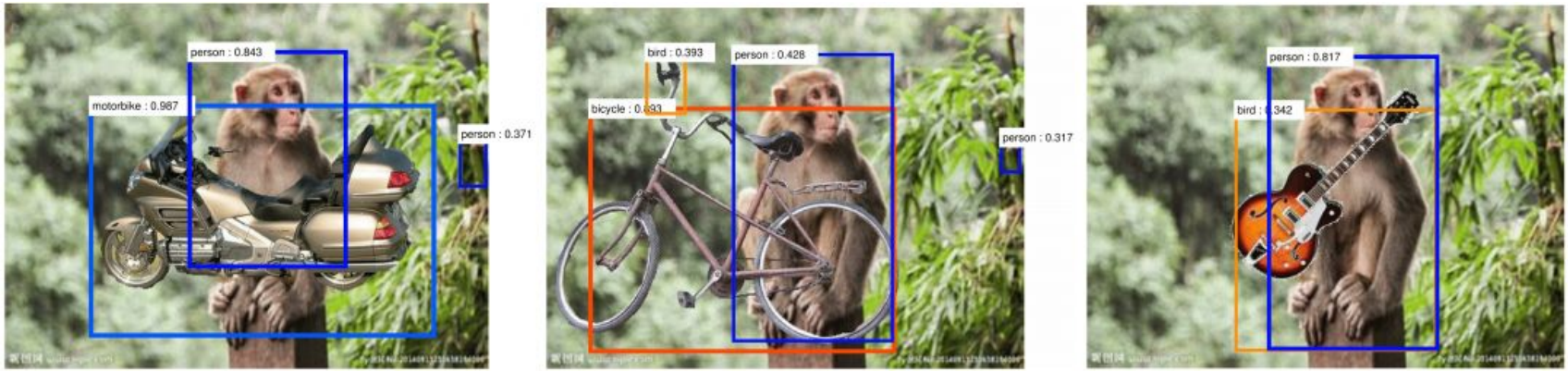
Image space dimensionality = W * H * 3 ~= (227 * 227 * 3 = 150K)

Training set covers only tiny subset of almost infinite number of all possible images. Adversarial examples lies in the space **far from learned manifold**.

Y. Bengio: The current incarnation of deep neural networks **have a tendency to learn surface statistical (superficial cues)** regularities as opposed to high level abstraction.



Adversarial examples - context level



Context plays an important role in recognition → **problem when test distribution != train distribution**

Neural nets **often overfits to typical context of the object**

Visual Concepts and Compositional Voting, J. Wang et al. 2017

“Adversarial” examples from real life

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

S. Beery et al., Recognition in Terra Incognita, 2018

Adversarial examples - segmentation and detection.

Figure 8: Figure taken from Xie et al. (2017). The top row is the input (adversarial perturbation already added) to the segmentation network, and the bottom row is the output. The red, blue and black regions are predicted as *airplane*, *bus* and *background*, respectively.

C. Xie et al., Adversarial Examples for Semantic Segmentation and Object Detection, The Johns Hopkins University, 2017

Robustness

Elevation \ Azimuth	90	135	180	225	270
0	-	0.713	0.769	0.930	0.319
30	0.900	1.000	0.588	1.000	0.710
60	0.255	0.100	0.148	0.296	0.649

Figure 2: Figure taken from Qiu and Yuille (2016). UnrealCV allows vision researchers to easily manipulate synthetic scenes, e.g. by changing the viewpoint of the sofa. We found that the Average Precision (AP) of Faster-RCNN (Ren et al., 2015) detection of the sofa varies from 0.1 to 1.0, showing extreme sensitivity to viewpoint. This is perhaps because the biases in the training cause Faster-RCNN to favor specific viewpoints.

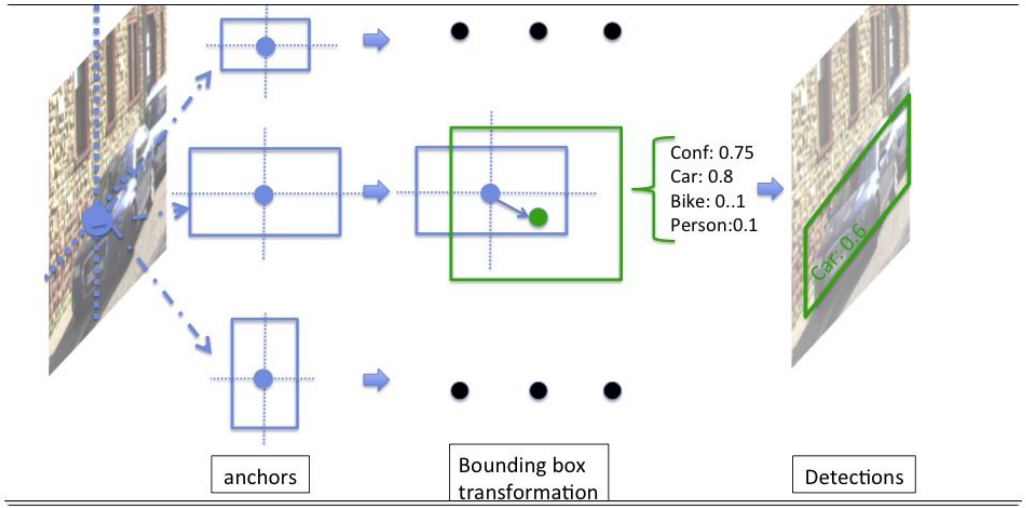
Robustness

Mask R-CNN

INZNAK

Traffic monitoring system, **for intelligent road sign** using multiple sensors. The sign will be equipped with different sensors ...

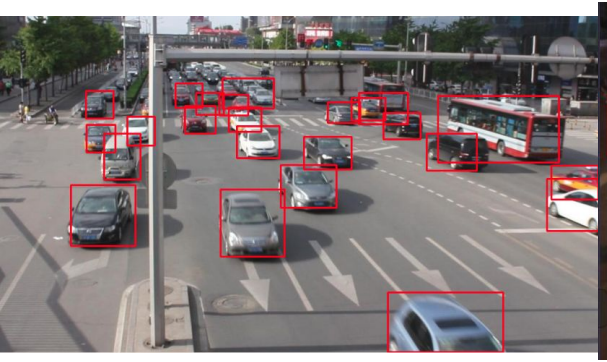
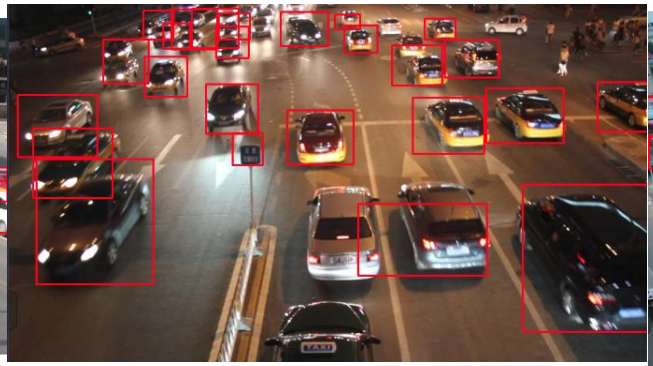
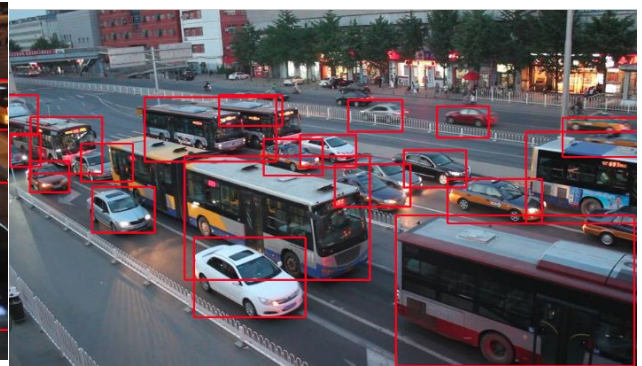
Computer Vision. The goal is to perform reliable vehicle counting / detection **without any manual calibration**. This means that system must be robust to various road conditions, **camera angles** and energy efficient.



Vehicle detection example.

UA-Detrac dataset. 140 thousand frames in the UA-DETRAC dataset and 8250 vehicles that are manually annotated, leading to a total of **1.21 million labeled bounding boxes of objects**, cameras at 24 different locations at Beijing and Tianjin.

No camera is shared between train and test dataset, however the dataset clearly show some similarities.

Vehicle detection (real life)

Vehicle detection in the wild

- Seemingly much simpler scenario
- Significant drop in results
- **Obtained results similar to background subtraction methods** (no training required)
- Deep nets can provide great improvement in some scenarios, and none in others.



Vehicle detection in the wild

Detector	Overall	Cars & Vans	Trucks
SqueezeDet	**42.74%**	**75.2%**	8.03%
Background subtraction	41.24%	48.47%	**31.8%**

- Problem with generalization to new viewpoints
- Test distribution of vehicles != train distribution of vehicles
- No generalization to new object subcategories
- DL algorithms needs to be more robust to be deployed to various environments

Learning for the long-tail

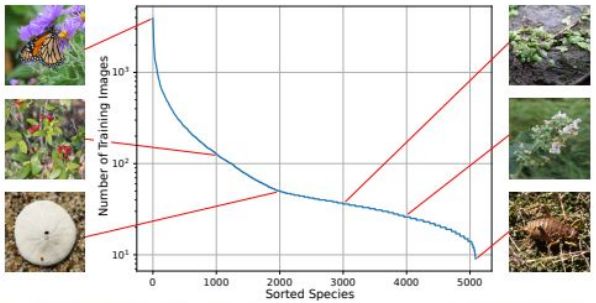
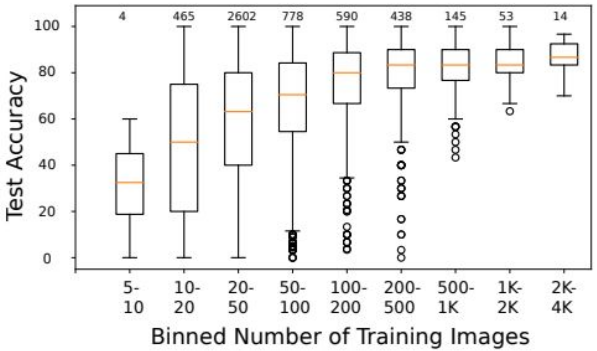
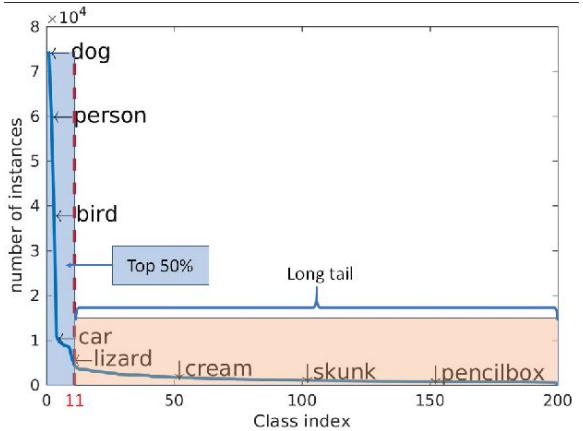


Figure 2. Distribution of training images per species. iNat2017 contains a large imbalance between classes, where the top 1% most populated classes contain over 16% of training images.





- Accuracy drops massively with smaller number of training examples
- Often examples with small number of training examples are the ones **we are the most interested in (anomaly detection, really big trucks, etc.)**

G. V. Horn et al., The iNaturalist Species Classification and Detection Dataset, 2017

Autonomous driving

- First death (Tesla 2016) - the white side of the tractor trailer against a brightly lit sky
- Uber death crash (2018):
  - too low threshold. Cyclist ignored as false positive
- Volvo plans postponed in 2017 by 4 years (also only level4 (fully automatic in certain conditions) instead of level5)

- **Model-free supervised learning -** does not scale when we want to reach near 100% accuracy
- Situations we are the most interested in occurs very rarely
- How to model occlusions (which number is infinite)?

Motivation

1. It is easy to see that a single object can b**e occluded in an exponential number of ways**
2. Humans are very adaptive to changes in context (...) by contrast, Deep Nets appear **more sensitive to context**

These complexity considerations mean that certain visual tasks require dealing with an **exponential number of hypotheses**. This is highly problematic from a machine learning perspective, because such algorithms may require, in principle, **exponential amounts of data**

In short, the standard vision evaluation methods **will start having problems as we develop increasingly complicated vision models.**

From an intuitive perspective, there will be many rare events which will not be well represented in the datasets.

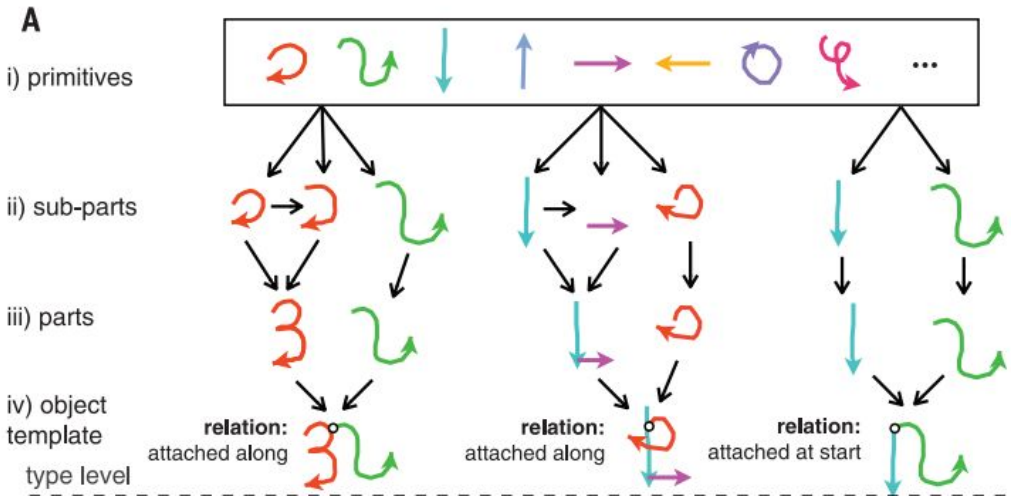
How to learn models which are exponentially complex when there is only limited amounts of data available?

Compositional models

Compositional Models represent objects in terms of:

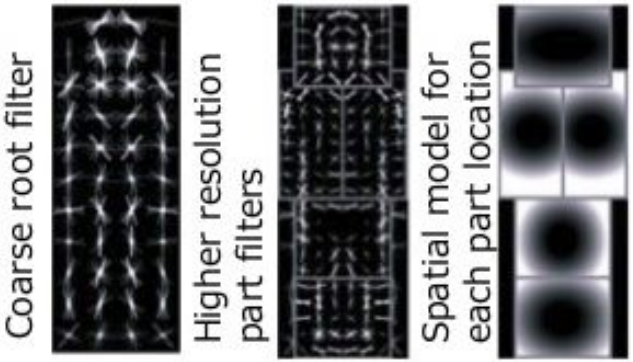
- object parts
- their spatial relations

Problem - what are the “words” for real images?



J. B. Tenenbaum et al., Human-level concept learning through probabilistic program induction, 2015

Deformable Part Model (DPM) [1, 2]



物体のモデル化
(1) Root filter: 大まかな形状
(2) Parts model:
Part filters(移動可能な部品)
+ Spatial models(移動コスト)

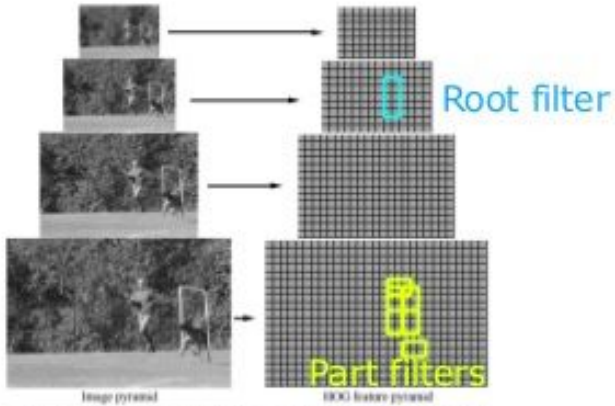


Figure 2. The HOG feature pyramid and an object hypothesis defined in terms of a placement of the root filter (near the top of the pyramid) and the part filters (near the bottom of the pyramid).

Semantic parts detection with visual concepts under occlusion

Our technique is based on the **hypothesis that semantic parts are represented by populations of neurons rather than by single filters.** We propose a clustering technique to extract part representations, which we call **Visual Concepts**.

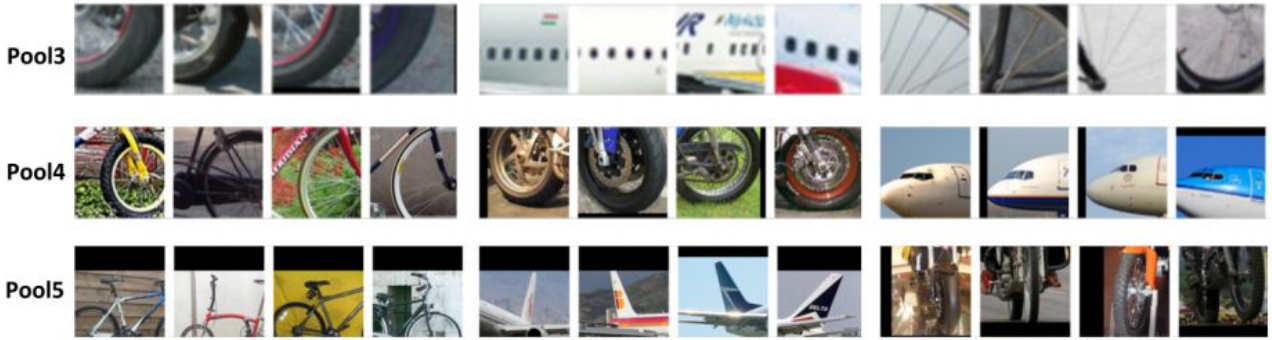


Figure 4: Figure taken from Wang et al. (2015). The visual concepts obtained by population encoding are visually tight and we can identify the parent object class pretty easily by just looking at the mid-level concepts.

Unsupervised learning of object semantic parts from internal states of CNNs by population encoding

Task - semantic part detection

Object: *car*; SP #17: *headlight*

VC #160

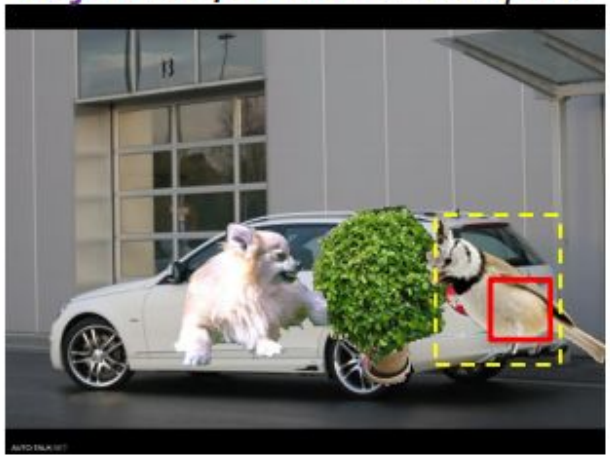
VC #245

VC #091

List of voted VC's:

1. **#160:** score = 0.393
   $(\Delta x, \Delta y) = (0,0)$
2. **#245:** score = 0.091
   $(\Delta x, \Delta y) = (+5,+1)$
3. **#091:** score = 0.053
   $(\Delta x, \Delta y) = (+6,+3)$

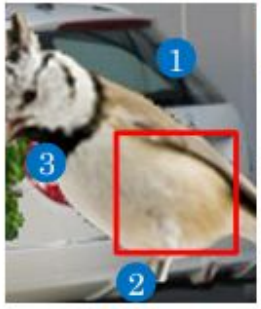
Object: *car*; SP #20: *licence plate*



VC #073

VC #235

VC #232



List of voted VC's:

1. **#073:** score = 0.020
   $(\Delta x, \Delta y) = (0,-6)$
2. **#235:** score = 0.012
   $(\Delta x, \Delta y) = (0,+4)$
3. **#232:** score = 0.007
   $(\Delta x, \Delta y) = (-5,-3)$

DeepVoting: DeepVoting: A Robust and Explainable Deep Network for Semantic Part Detection under Partial Occlusion

Category	No Occlusions						L1				L2				L3			
	KVC	**DVC**	**VT**	**FR**	**DV**	**DV+**	**VT**	**FR**	**DV**	**DV+**	**VT**	**FR**	**DV**	**DV+**	**VT**	**FR**	**DV**	**DV+**
airplane	15.8	26.6	30.6	56.9	59.0	**60.2**	23.2	35.4	**40.6**	**40.6**	19.3	27.0	31.4	**32.3**	15.1	20.1	**25.9**	25.4
bicycle	58.0	52.3	77.8	90.6	89.8	**90.8**	71.7	77.0	83.5	**85.2**	66.3	62.0	78.7	**79.6**	54.3	41.1	**63.0**	62.5
bus	23.8	25.1	58.1	**86.3**	78.4	81.3	31.3	55.5	56.9	**65.8**	19.3	40.1	44.1	**54.6**	9.5	25.8	30.8	**40.5**
car	25.2	36.5	63.4	83.9	80.4	**80.6**	35.9	48.8	56.1	**57.3**	23.6	30.9	40.0	**41.7**	13.8	19.8	27.3	**29.4**
motorbike	32.7	29.2	53.4	63.7	65.2	**69.7**	44.1	42.2	51.7	**55.5**	34.7	32.4	41.4	**43.4**	24.1	20.1	29.4	**31.2**
train	12.3	12.8	35.5	59.9	59.4	**61.2**	21.7	30.6	33.6	**43.7**	8.4	17.7	19.8	**29.8**	3.7	10.9	13.3	**22.2**
mean	28.0	30.4	53.1	73.6	72.0	**74.0**	38.0	48.3	53.7	**58.0**	28.6	35.0	42.6	**46.9**	20.1	23.0	31.6	**35.2**

Semantic parts detection better and Faster R-CNN on heavy occluded objects.

No occlusions at all in training set.

Conclusions

- Deep Net performance on benchmarked datasets, no matter how large, may f**ail to extend to good performance** images outside the dataset.
- Context plays a crucial role in object detection
- Lots of work happening right now with minimal supervision (self-supervised learning)
- Compositional models are gaining some attention (i.e. fine-grained classification)

Some reading

Compositional models:

D. George et al., A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs, **Vicarious**, 2017

J. B. Tenenbaum et al., Human-level concept learning through probabilistic program induction, 2015

Deep Nets understanding:

J. Jo, Y. Bengio, Measuring the tendency of CNNs to Learn Surface Statistical Regularities, 2017

Why does deep and cheap learning work so well?
https://blog.acolyer.org/2016/10/05/why-does-deep-and-cheap-learning-work-so-well/

General AI:

J. B. Tenenbaum et al., Building Machines That Learn and Think Like People, 2016

G. Marcus, Innateness, AlphaZero, and Artificial Intelligence , 2018