Machine Learning for Vision

Winter 2016

Roland Memisevic

Lecture 0, January 6, 2016

IFT 6268, Winter 2016

- Classes:
  - Tuesdays 2:30pm-4:30pm H-260 Pav. Claire-McNicoll
  - Wednesdays 3:30pm-5:30pm Z-205 Pav. Claire-McNicoll
- Instructor:

  Roland Memisevic,
  3349, Pav. Andre-Aisenstadt
- Office hours:

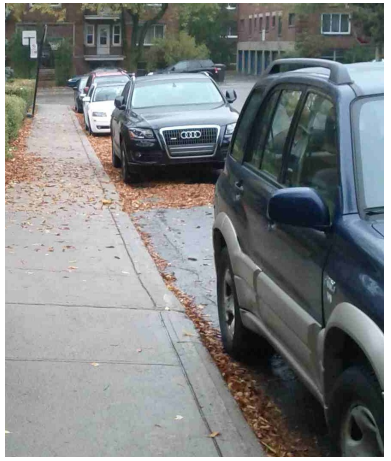
  drop in or by appointment
- Course website:

  http://www.iro.umontreal.ca/~memisevr/teaching/ift6268_2016/index.html

Objectives

- Learn about the recent advances in **data driven vision.**
- Learn how to apply some **state-of-the-art learning and inference techniques in vision tasks.**
- Learn about the basics and peculiarities of **natural images statistics.**
- (+ Get some ideas about visual information processing in biological systems.)

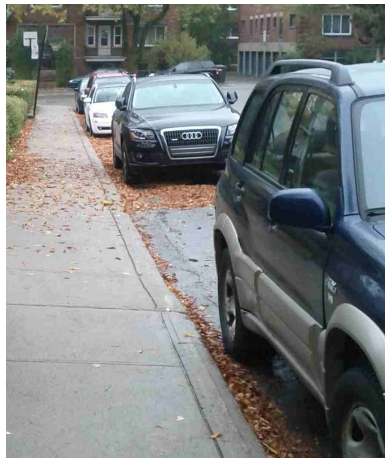
What this course is about



how many cars in the picture?

- Vision looks easy to humans.
- It is robust and flexible.
- It runs on fairly general-purpose hardware.

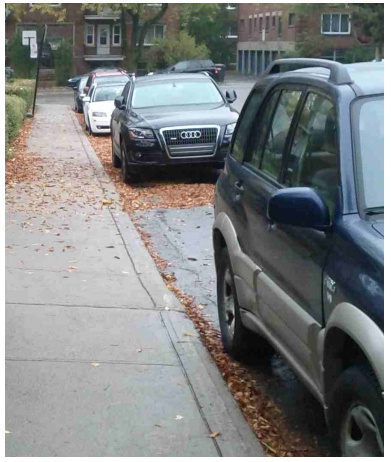
What this course is about



how many cars in the picture?

- Computer vision spent $\approx$ 50 years trying to mimic human vision.
- Huge inventory of tools: edge detectors, corner detectors, descriptors (eg. SIFT), optic flow, hough transform, projective geometry...
- Unfortunately, it is difficult to make these work nicely together.

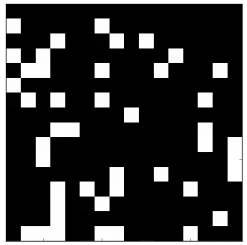
What this course is about



how many cars in the picture?

- Huge progress in recent years, based on a single simple idea:
- *Images are not random*. $\rightarrow$ Treat vision as a *statistical inference* task.

A lower bound on the number of all images



- Assume your retina was only 16×16 pixels large and you could see only black and white.
- There are still $2^{16*16} = 2^{256}$ *possible* images.

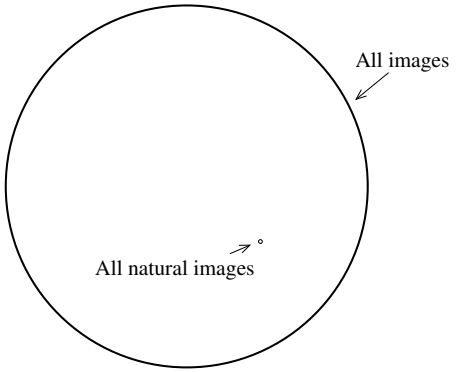
(=115792089237316195423570985008687907853269984665640564039457584007913129639936)

- So there are more tiny binary images than there are atoms in the universe.
- And even more large color images.

An upper bound on the number of images you will see in your life

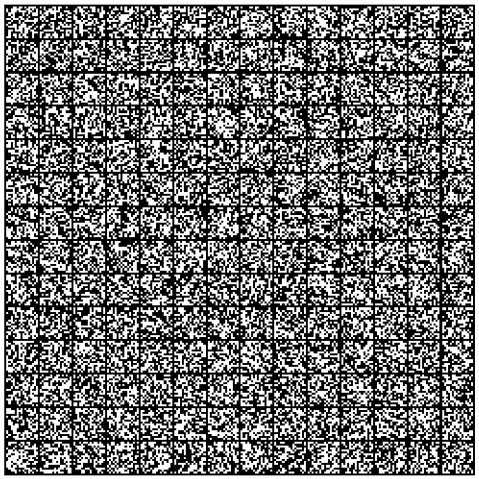
- Assume you see 100 images per second, 3600 seconds per hour, 24 hours per day.
- This is $<$ 10 mio images per day, or 3.65 billion images per year.
- So you will see $<$ 300 billion images in your life and you had seen $<$ 10 billion images when you turned 3.
- This number is tiny compared to the number of *possible* images.
- Yet, at that age you were a champion at recognizing and reasoning about unfamiliar objects.
- The number of *labeled* images is much much smaller yet.

Natural images are not random

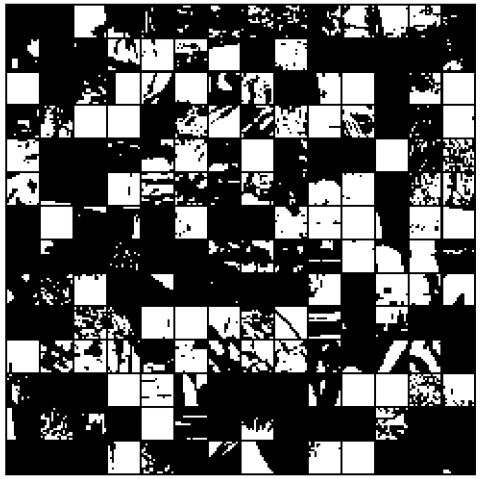


- As compared to the number of possible images, there is a diminishingly small number of *natural* images!

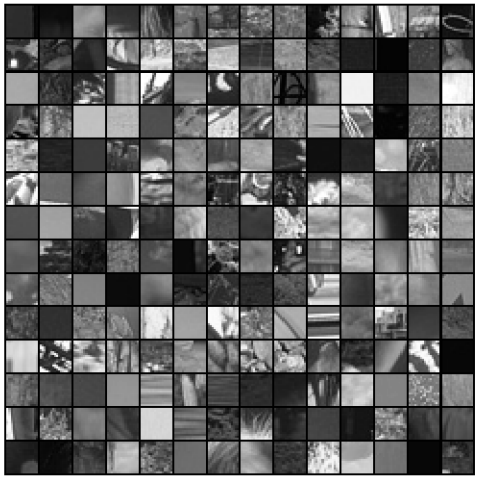
Random images



Natural images (berkeley database)



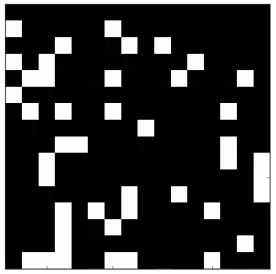
Natural images (grayscale)



View from information theory

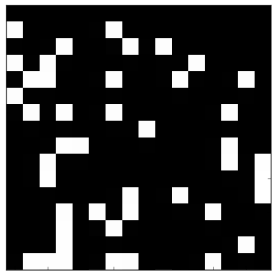
- The distribution over natural images has *low entropy*.

View from information theory



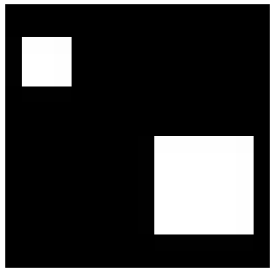
- How many bits will you need to transmit (or save) this image?

View from information theory



- If images are “random”, you will need 256 bits on average to transmit each.

View from information theory



- If your images are structured, you will need much fewer bits.
- For example, what if the images contain two square blocks of random size at random locations?
- (Hyvarinen et al, 2009)

View from information theory

- You can transmit the upper-left corner and the bottom-right corner each with 8 bits (4 for the vertical, 4 for the horizontal direction), making it $2 \times 16 = 32$ bits for both squares.
- (It could be more efficient than that.)

View from information theory

- Caveat: Neural codes, ironically, are very *high*-dimensional. It is the entropy of each individual code element that is small. This leads to *sparse* representations.

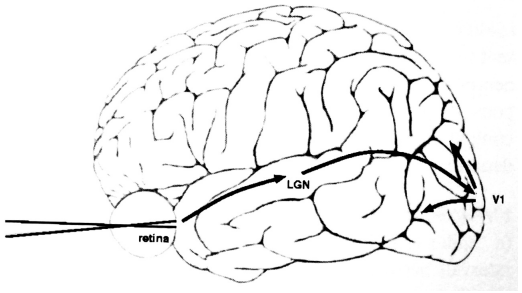
View from statistics

- Another way to state that the information content is small is to say that there are *dependencies* among the pixels.

View from statistics

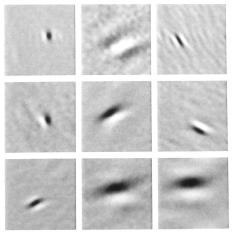
- A common way to reduce the dependencies is *Independent Components Analysis* (ICA)

View from neuroscience



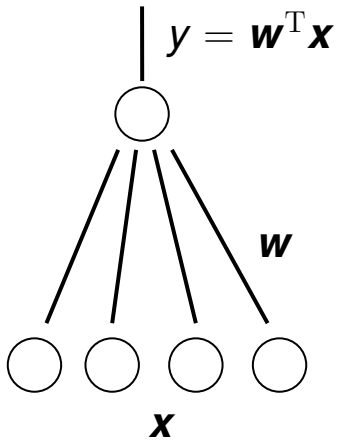
- Attneave 1954, Barlow 1961

What visual neurons like to see

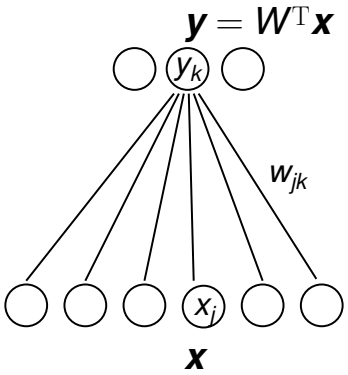


- Hubel and Wiesel, 1959

A very simple neuron abstraction



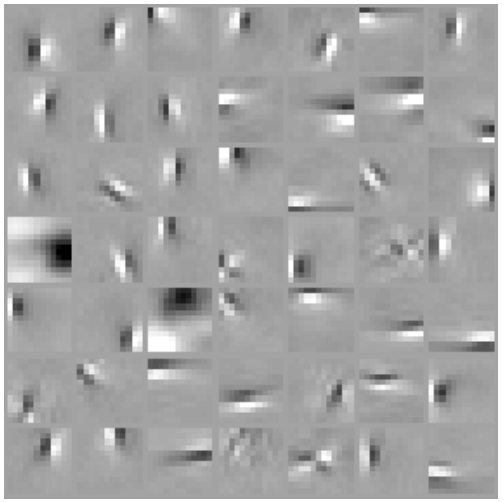
Two layers of neurons



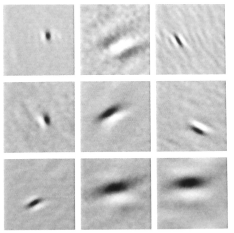
Learning criteria

- maximize independence (ica)
- minimize entropy (information theory)
- maximize sparseness (sparse coding)
- maximize probability of the data (eg. boltzmann machines, mixture models)
- learn to reconstruct from bottleneck (autoencoders, kmeans)
- **supervised learning** (eg. learn to classify objects)

Learned receptive fields

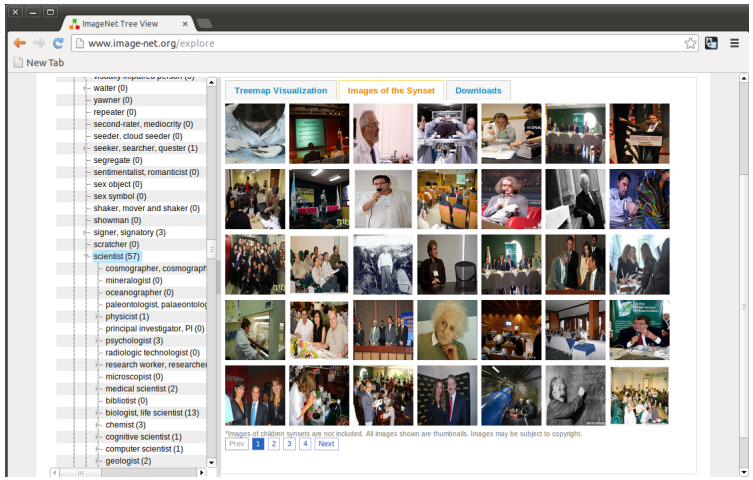


learned receptive fields

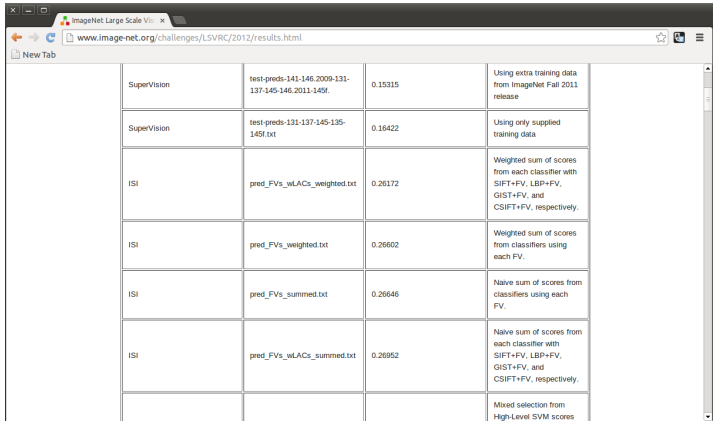


real receptive fields

ImageNet challenge

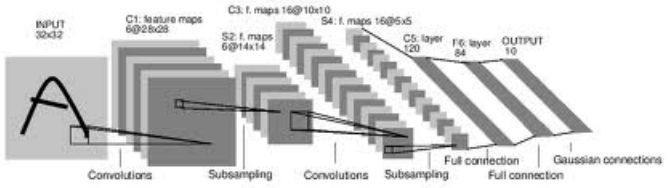


ImageNet challenge



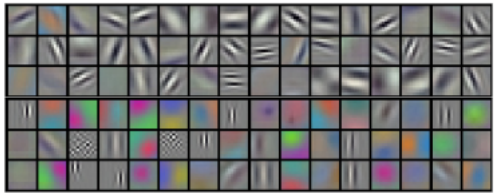
- Krizhevsky, et al. 2012

Convolutional networks



- LeCun et al. 1998
- Fukushima 1980 (without learning)

Low-level features



- Krizhevsky, et al. 2012

High-level features

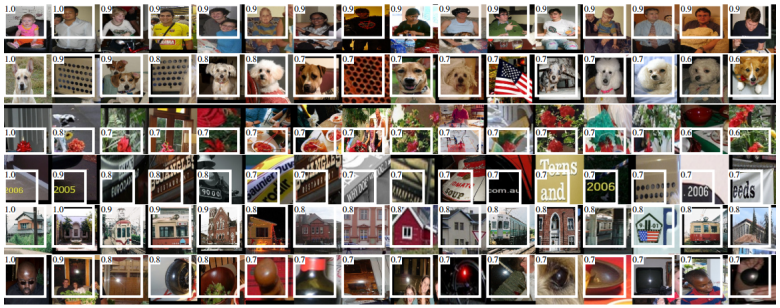
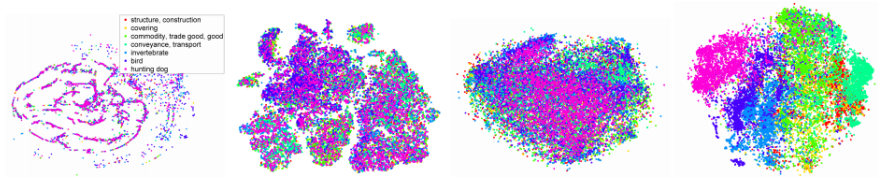


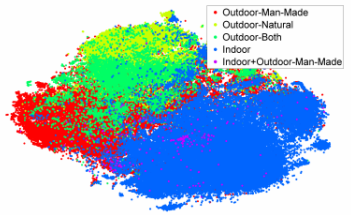
Figure 4: Top regions for six pool$_5$ units. Receptive fields and activation values are drawn in white. Some units are aligned to concepts, such as people (row 1) or text (4). Other units capture texture and material properties, such as dot arrays (2) and specular reflections (6).

Girshick, Donahue, Darrell, Malik (!); 2014

Convnet features for generic recognition



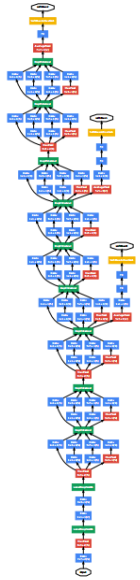
(a) LLC (b) GIST (c) DeCAF$_1$ (d) DeCAF$_6$



non-imagenet classes:

(Donahue et al, 2013)

GoogLeNet



- exercise in (a) scaling up, (b) unconventional neurons/architectures
- wins ImageNet 2014 with **6.66%** top-5 error rate
- vision solved?

Vision is more than object recognition

how many cars in the picture?

Vision is more than object recognition

how many cars in the picture?

There are things images can't teach you

There are things images can't teach you

There are things images can't teach you

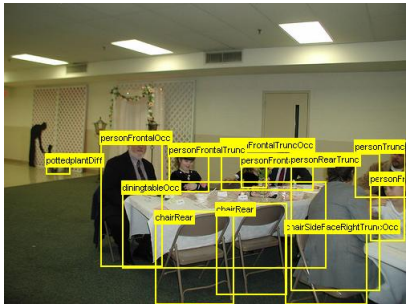
There are things images can't teach you

There are things (still) images can't teach you

how many chairs in the picture?

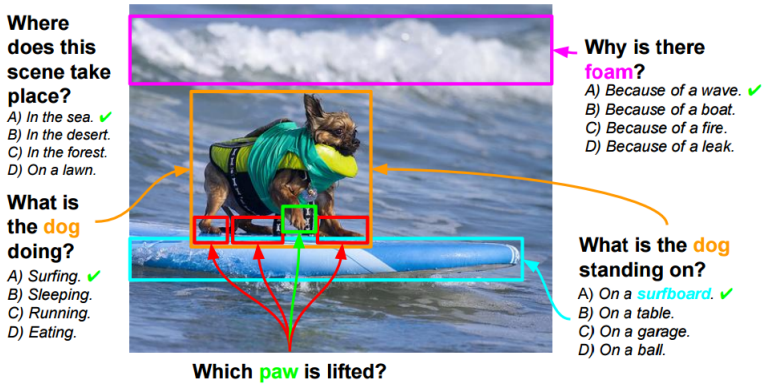
(Buelthoff and Buelthoff, 2003)

There are things (still) images can't teach you

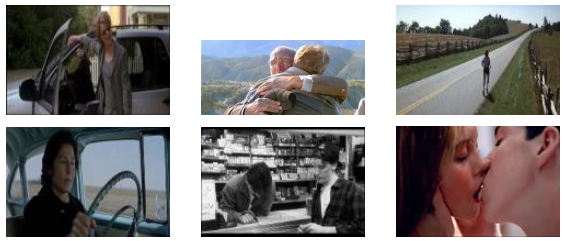


how many chairs in the picture?

Zhu, Groth, Bernstein, Li (2015)



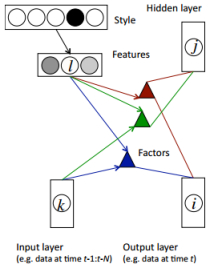
Activity recognition example



("Hollywood 2", Marszałek et al., 2009)

- Convolutional GBM (Taylor et al., 2010)
- hierarchical ISA (Le, et al., 2011)

Tracking



(Taylor, et al.; 2010)

Major conferences and journals

- **ICLR:** International Conference on Learning Representations
- **NIPS:** Neural Information Processing Systems
- **CVPR:** International Conference on Computer Vision and Pattern Recognition
- **ICCV:** International Conference on Computer Vision
- **ICML:** International Conference on Machine Learning
- **ECCV:** European Conference on Computer Vision
- **PAMI:** IEEE Transactions on Pattern Analysis and Machine Intelligence
- **Neural Computation**
- **JMLR:** Journal of Machine Learning Research

Course outline

1. Fourier representations and Gabor features, image statistics, visual features
2. Aspects of biological vision
3. Supervised learning
4. Convolutional network, types, variations, tricks
5. Advanced topics: Attention, vision and language, vision and robotics

Learning approach

- Readings will be posted and should be read before each class.
- Lectures will explain and motivate the concepts with real world examples.
- Student presentations of recent papers to discuss recent/novel/speculative/applied ideas.
- Several hands-on assignments to get an idea for how the methods work on actual data.
- Final projects are research based. Eg. evaluation/comparison of an approach from a recent paper, prototype/discussion of a new idea or variation of an existing one.

Marking scheme

- readings (10 %)
- participation in class (20 %)
- assignments (30 %)
- term project (40 %)

Relation to other courses and areas

- **Image Processing, Computer Vision:** Focus on *data* and *learning* (and *bio-inspired* as a consequence).
- **Neuroscience:** The brain (and neuroscience) is utterly complex and detailed. We will abstract away *a lot* of these details.
- **Machine Learning:** Images have *strong structure*. Black-box classifiers (like SVM) and fully Bayesian / variational methods not always the best choice.
- **Deep Learning:** Focus on images (i) to solve vision tasks, (ii) to study the models.