
Information
Signal
Image
viSion



Journée GdR ISIS TRECVid/Deep
20 mars 2015

# Deep Learning Introduction

Matthieu Cord

LIP6/UPMC

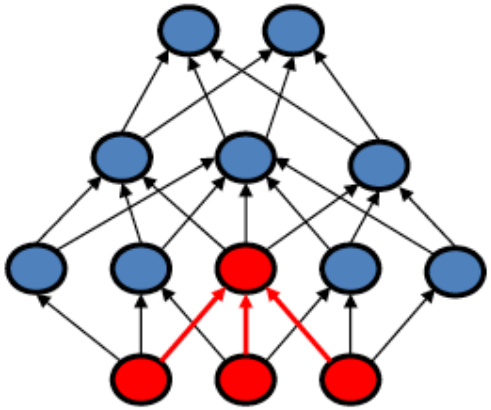
# Outline

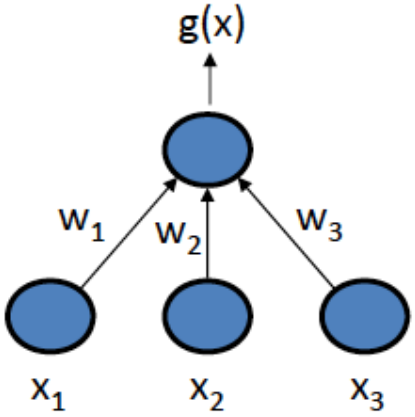
1. Key dates in deep learning
2. Deep learning for object recognition
3. Discussion

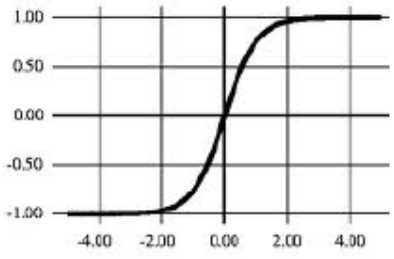
Neural network
Back propagation

*Nature*

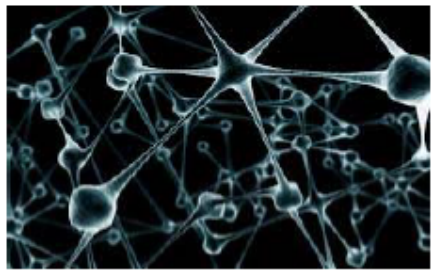
1986





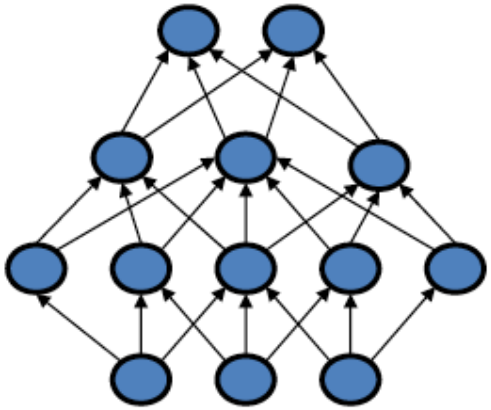

$$g(\mathbf{x}) = f(\sum_{i=1}^{d} x_i w_i + w_0) = f(\mathbf{w}^t \mathbf{x})$$



Neural network
Back propagation

*Nature*

1986



- Solve general learning problems
- Tied with biological system

But it is given up…

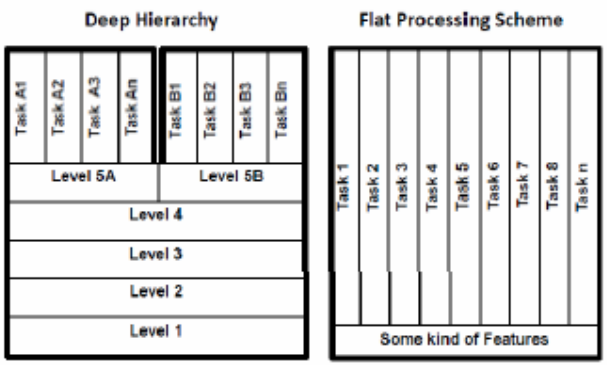
- Hard to train
- Insufficient computational resources
- Small training sets
- Does not work well

Neural network
Back propagation

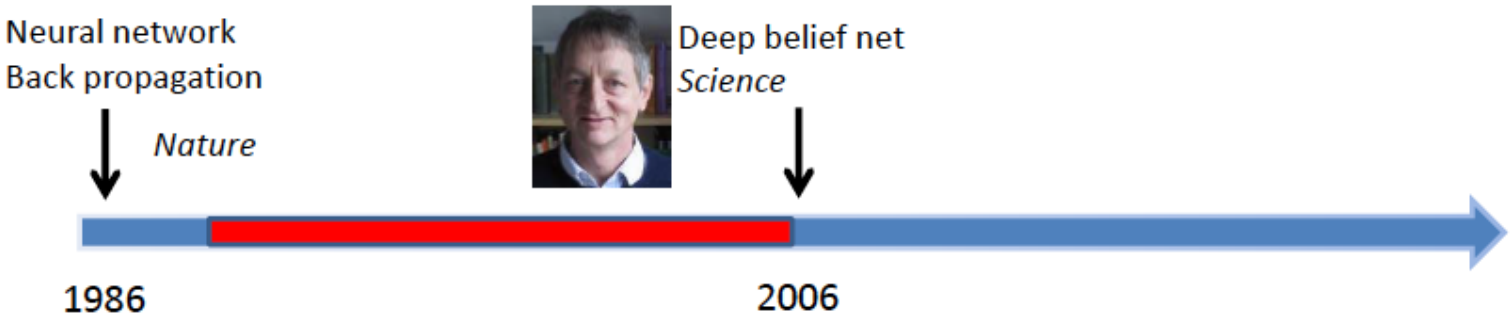
*Nature*

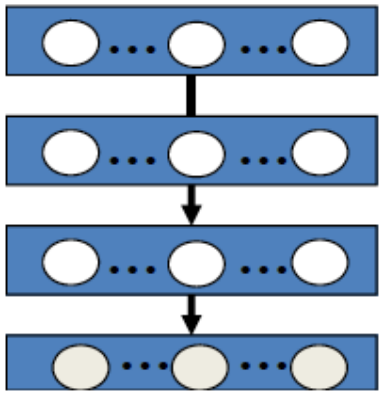

- SVM
- Boosting
- Decision tree
- KNN
- ...

- Flat structures
- Loose tie with biological systems
- Specific methods for specific tasks
  - Hand crafted features (GMM-HMM, SIFT, LBP, HOG)




Kruger et al. TPAMI'13

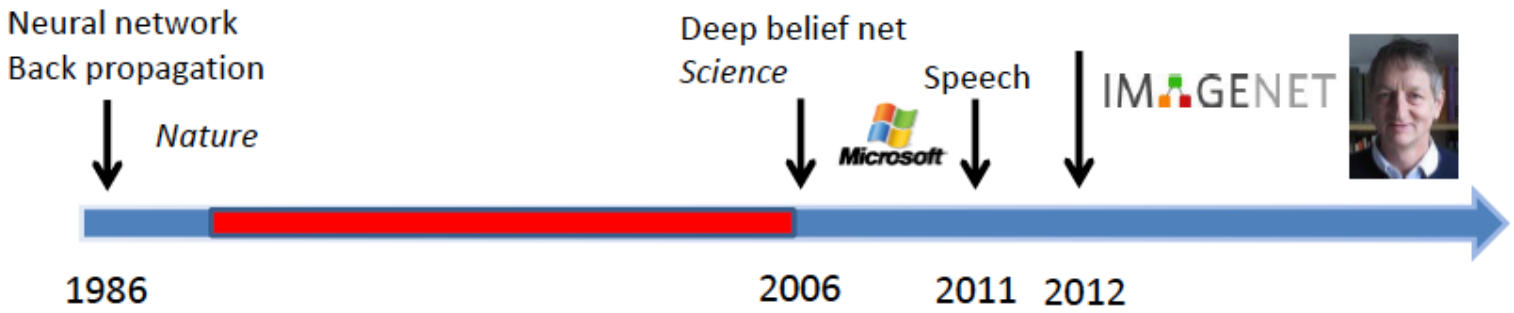


- Unsupervised & Layer-wised pre-training
- Better designs for modeling and training (normalization, nonlinearity, dropout)
- New development of computer architectures
  - GPU
  - Multi-core computer systems
- Large scale databases

**Big Data !**

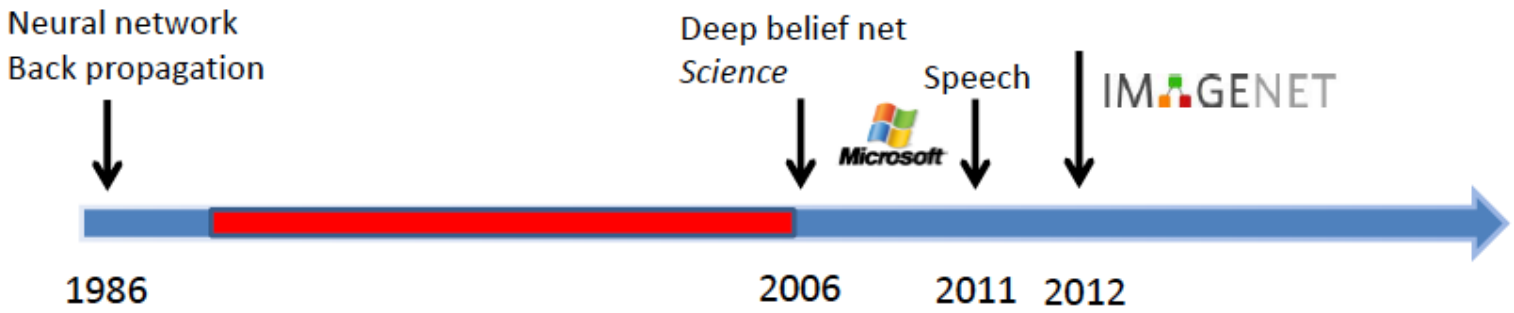
@Xiaogang Wang, CUHK, ICIP Tutorial 2014




| Rank | Name | Error rate | Description |
|---|---|---|---|
| 1 | **U. Toronto** | 0.15315 | Deep learning |
| 2 | U. Tokyo | 0.26172 | Hand-crafted features and learning models. Bottleneck. |
| 3 | U. Oxford | 0.26979 | |
| 4 | Xerox/INRIA | 0.27058 | |

Object recognition over 1,000,000 images and 1,000 categories (2 GPU)

A. Krizhevsky, L. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," NIPS, 2012.



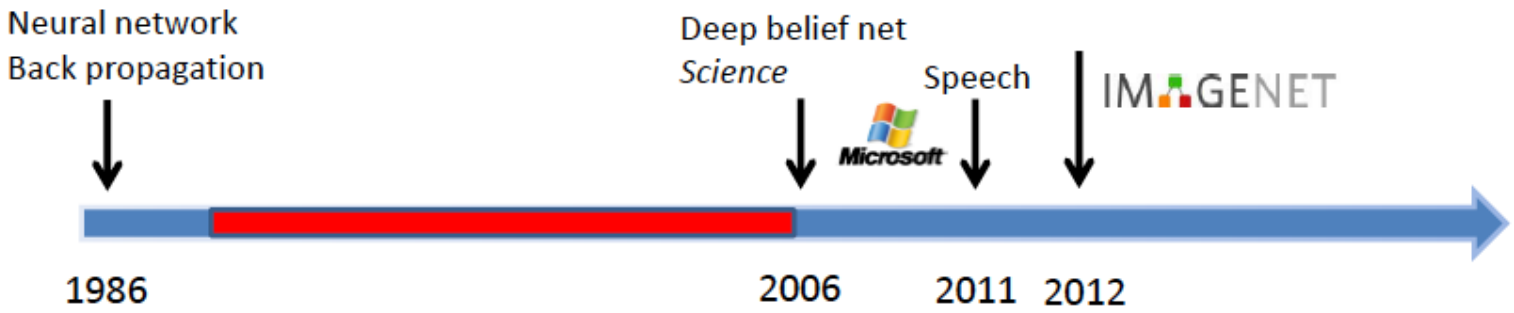

- ImageNet 2013 – image classification challenge

| Rank | Name | Error rate | Description |
|---|---|---|---|
| 1 | NYU | 0.11197 | Deep learning |
| 2 | NUS | 0.12535 | Deep learning |
| 3 | Oxford | 0.13555 | Deep learning |

MSRA, IBM, Adobe, NEC, Clarifai, Berkley, U. Tokyo, UCLA, UIUC, Toronto …. Top 20 groups all used deep learning

- ImageNet 2013 – object detection challenge

| Rank | Name | Mean Average Precision | Description |
|---|---|---|---|
| 1 | UvA-Euvision | 0.22581 | Hand-crafted features |
| 2 | NEC-MU | 0.20895 | Hand-crafted features |
| 3 | NYU | 0.19400 | Deep learning |



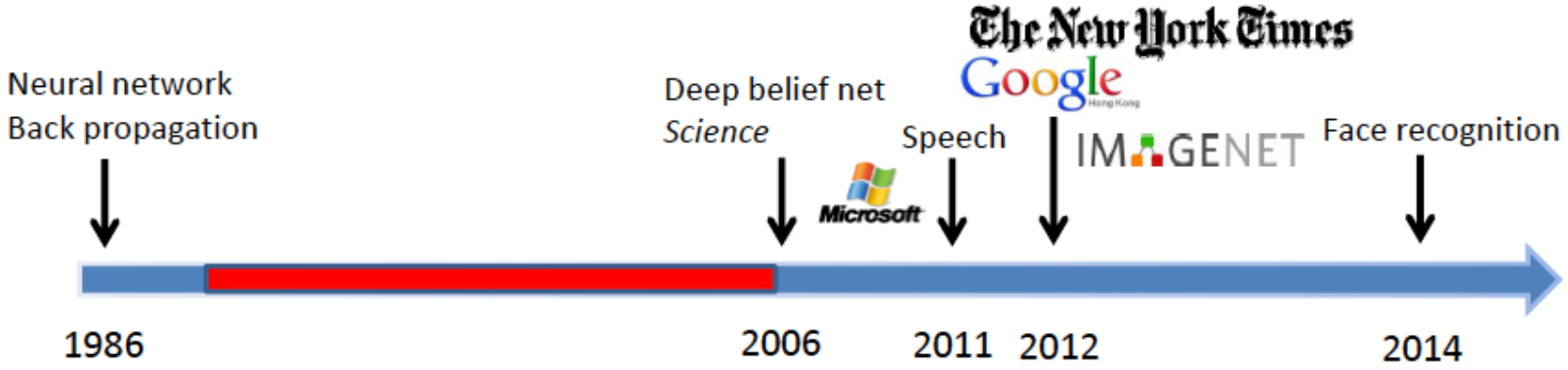

- ImageNet 2014 – Image classification challenge

| Rank | Name | Error rate | Description |
|---|---|---|---|
| 1 | Google | 0.06656 | Deep learning |
| 2 | Oxford | 0.07325 | Deep learning |
| 3 | MSRA | 0.08062 | Deep learning |

- ImageNet 2014 – object detection challenge

| Rank | Name | Mean Average Precision | Description |
|---|---|---|---|
| 1 | Google | 0.43933 | Deep learning |
| 2 | CUHK | 0.40656 | Deep learning |
| 3 | DeepInsight | 0.40452 | Deep learning |
| 4 | UvA-Euvision | 0.35421 | Deep learning |
| 5 | Berkley Vision | 0.34521 | Deep learning |

@Xiaogang Wang, CUHK, ICIP Tutorial 2014




- Deep learning achieves 99.15% face verification accuracy on Labeled Faces in the Wild (LFW), close to human performance

Y. Sun, X. Wang, and X. Tang. Deep Learning Face Representation by Joint Identification-Verification. NIPS, 2014.

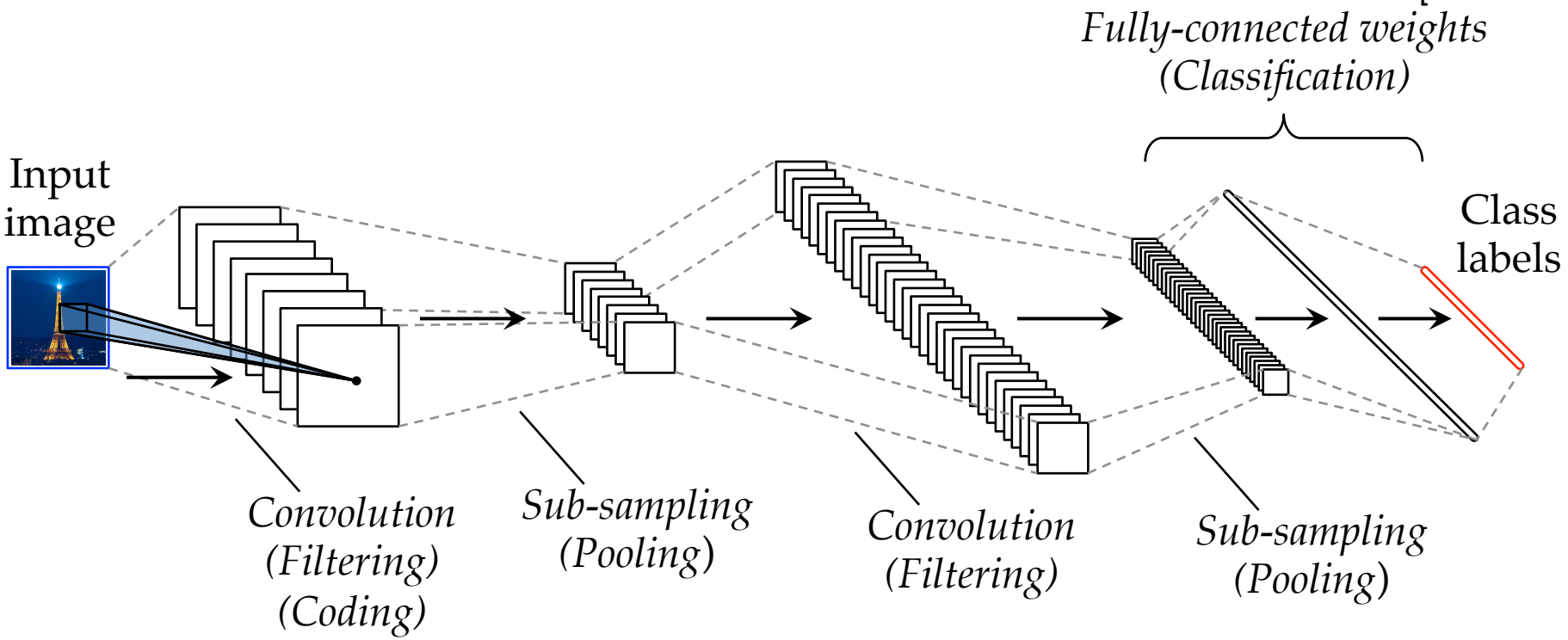
… and of course TRECVID 2014 Best results use deep learning!

# Outline

1. Key dates in deep learning
2. **Deep learning for object recognition**
   - Architecture
   - Results
   - Learning
3. Discussion

# Deep Convolutional Neural Networks (CNN)

[LeCun-89]

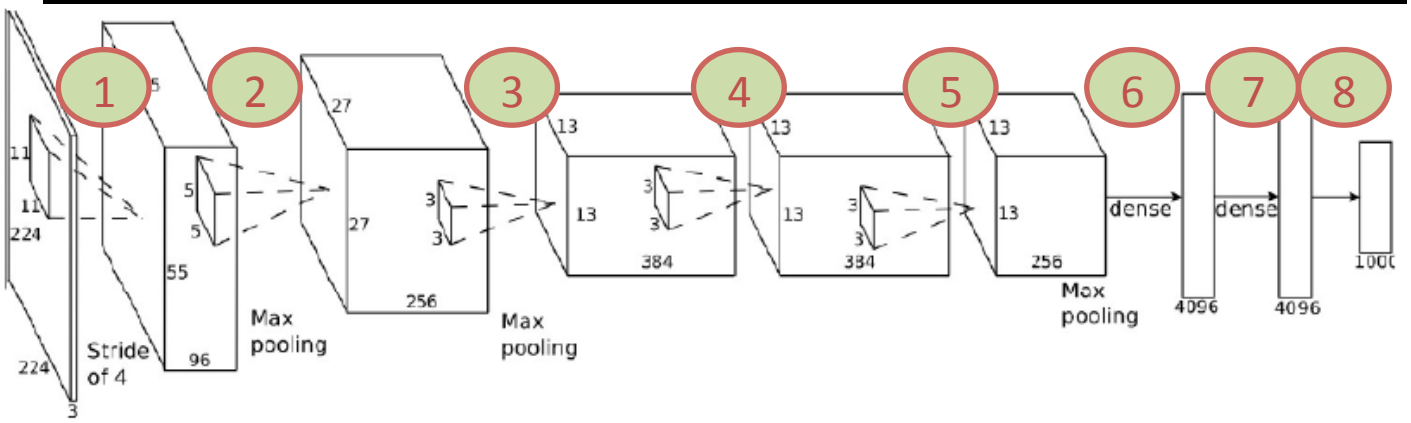



- **Convolution** uses local weights shared across the whole image
- **Pooling** shrinks the spatial dimensions
- Many other Deep Models (not convolutional):
  - Deep belief Net – Hinton'06 – Stack RBM
  - Auto-Encoder – Hinton and Salakhutdinov 06

# Large CNN [Slides @Fergus tutorial NIPS 2013]
# Architecture of the IMAGENET Challenge 2012 Winner:

## Krizhevsky et al. [NIPS2012]

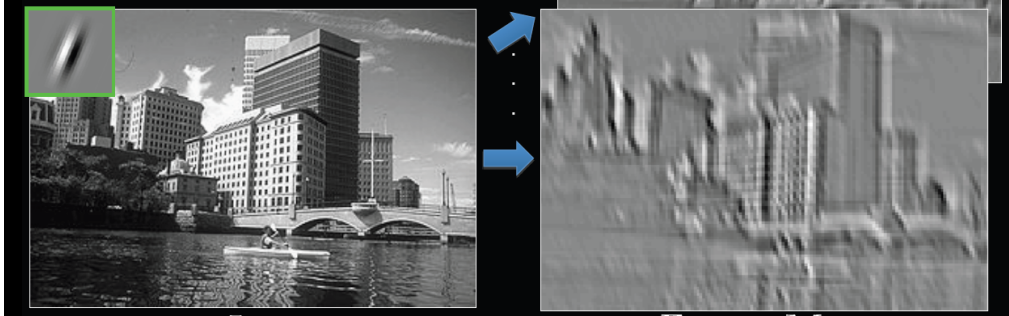
- Same model as LeCun'98 but:
  - Bigger model (8 layers)
  - More data ($10^6$ vs $10^3$ images)
  - GPU implementation (50x speedup over CPU)
  - Better regularization (DropOut)




- 7 hidden layers, 650,000 neurons, 60,000,000 parameters
- Trained on 2 GPUs for a week

## Filtering

- Convolutional
  - Dependencies are local
  - Translation equivariance
  - Tied filter weights (few params)
  - Stride 1,2,… (faster, less mem.)

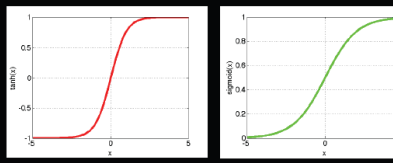


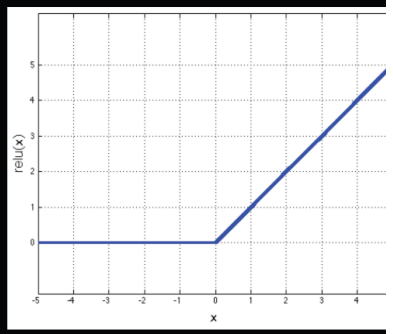
Input Feature Map

## Non-Linearity

- Non-linearity
  - Per-feature independent
  - Tanh
  - Sigmoid: 1/(1+exp(-x))
  - Rectified linear
    - Simplifies backprop
    - Makes learning faster
    - Avoids saturation issues

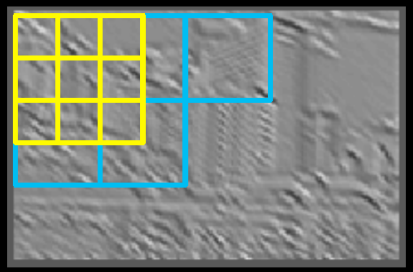
→ Preferred option



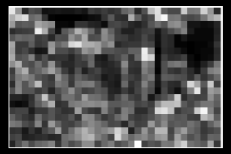



## Pooling

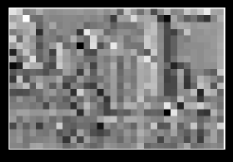
- Spatial Pooling
  - Non-overlapping / overlapping regions
  - Sum or max
  - Boureau et al. ICML'10 for theoretical analysis



Max



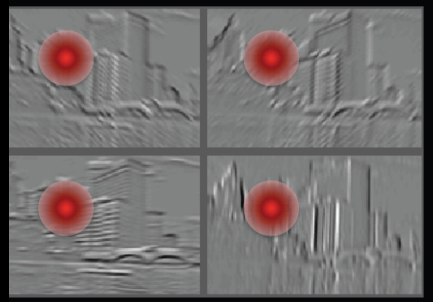
Sum



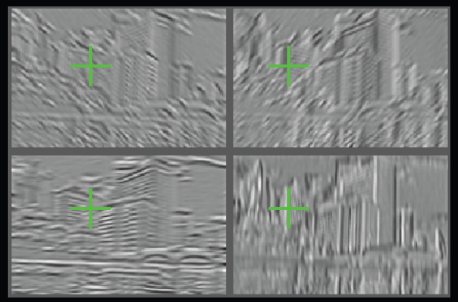
## Normalization

Contrast normalization (between/across feature maps)

- Local mean = 0, local std. = 1, "Local" → 7x7 Gaussian
- Equalizes the features maps

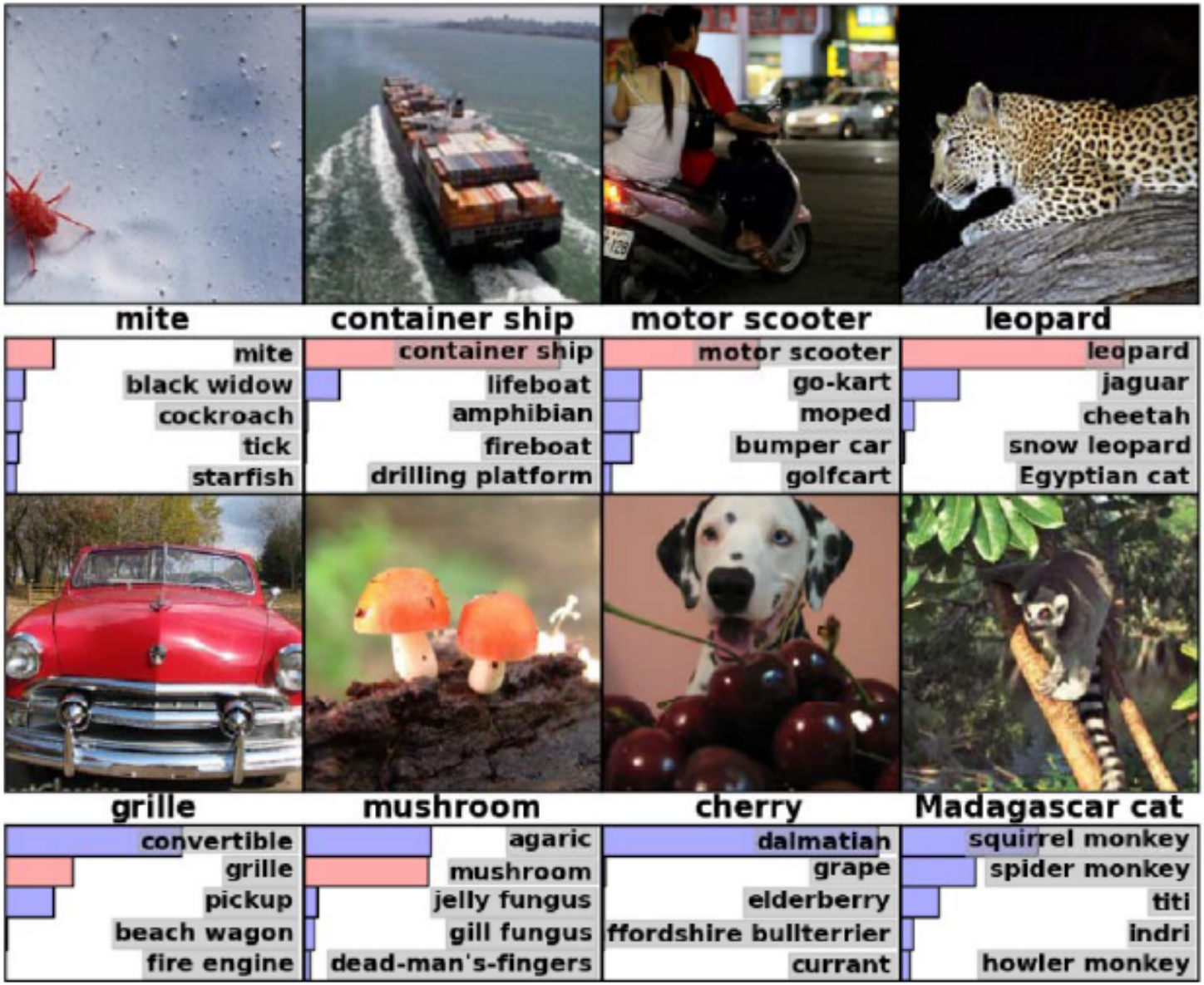


Feature Maps

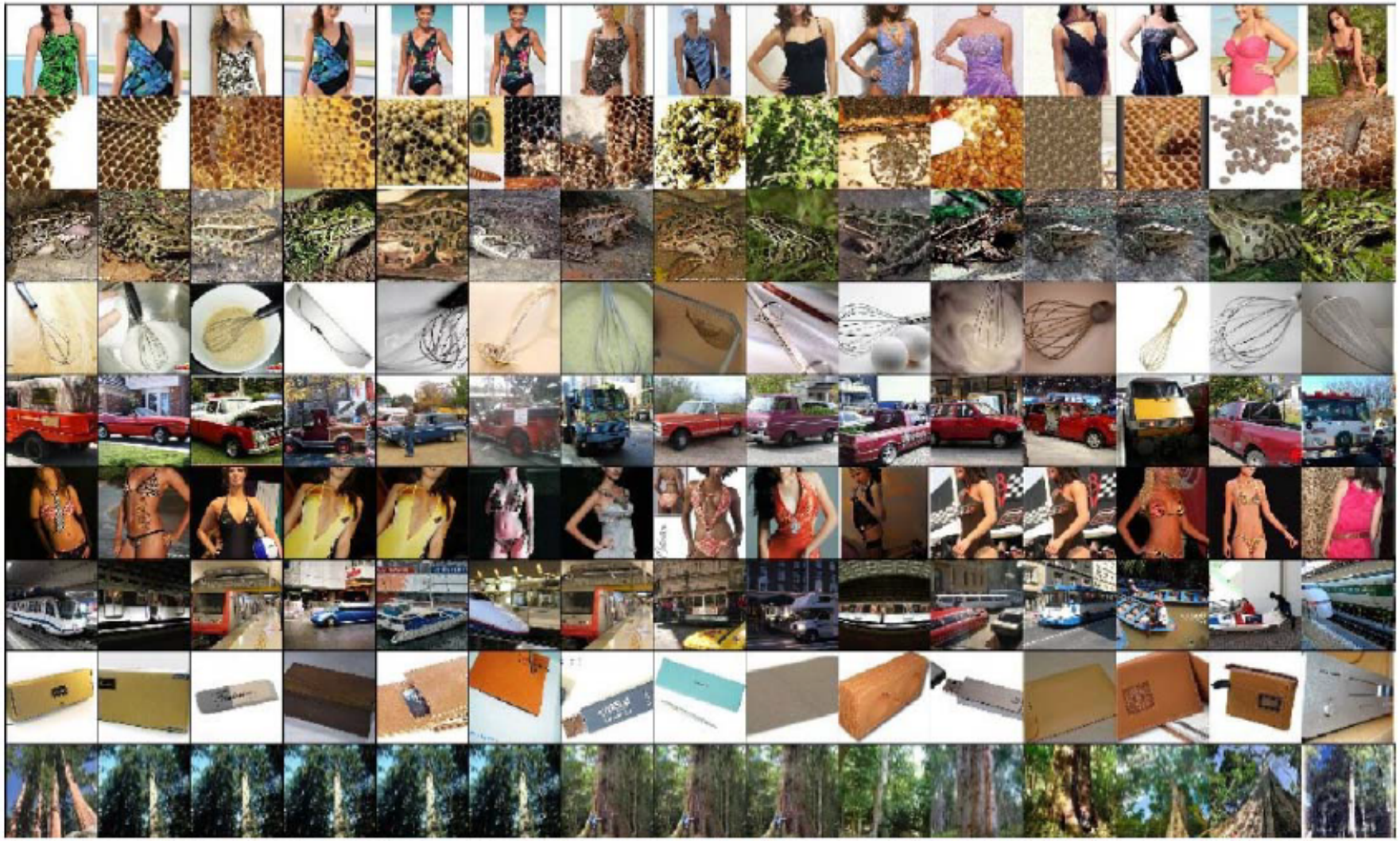


Feature Maps
After Contrast Normalization

# Image classification result



# Top hidden layer can be used as feature for retrieval



# Learning the deep CNN 2012

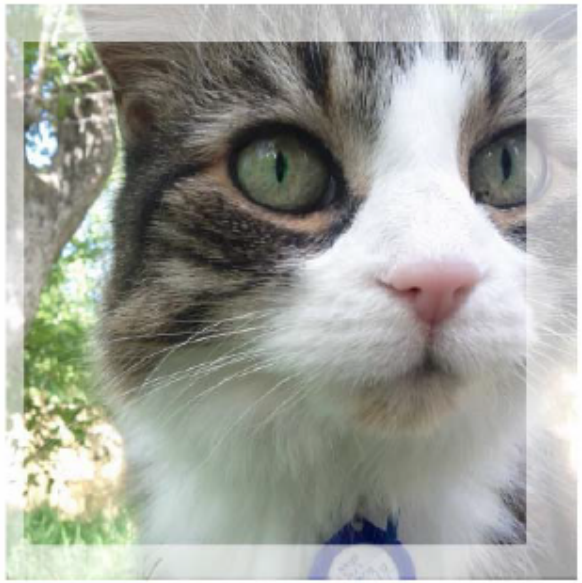
- Basics:
  - Backprop
  - Cross Validation
  - Grid search
- "New"
  - Huge computational resources (GPU)
  - Huge training set (1 million images)
  - Data augmentation - Pre-processing
  - Dropout

# Learning the deep CNN 2012

## Data Augmentation

- The neural net has 60M parameters and it overfits
- Image regions are randomly cropped with shift; their horizontal reflections are also included

Crop, flip
In train
AND
In test

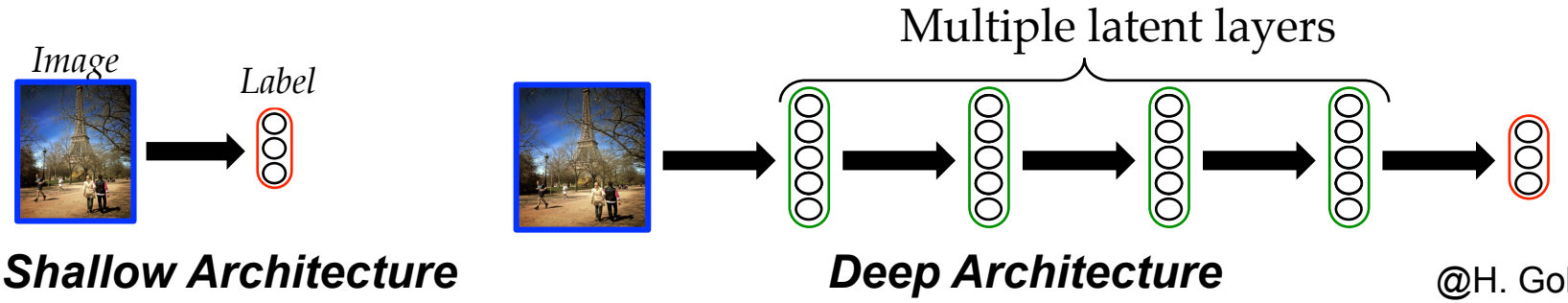


Krizhevsky 2012

Outline

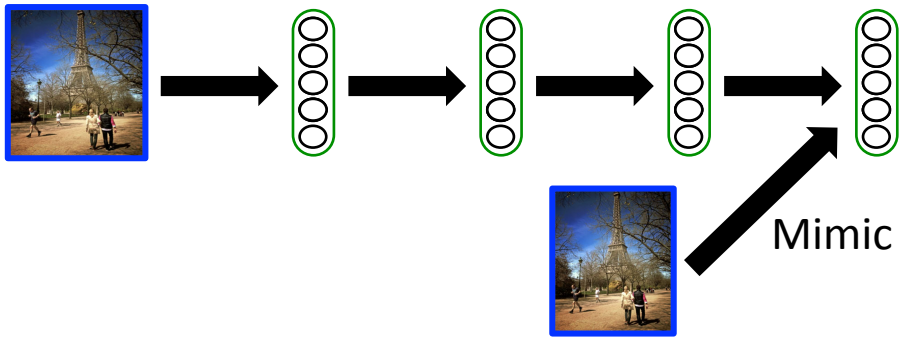
1. Key dates in deep learning
2. Deep learning for object recognition
3. **Discussion**
   - Deep vs Shallow (Why deep?)
   - Feature Learning vs Feature Engineering
   - Using deep in Computer Vision

# Deep vs shallow

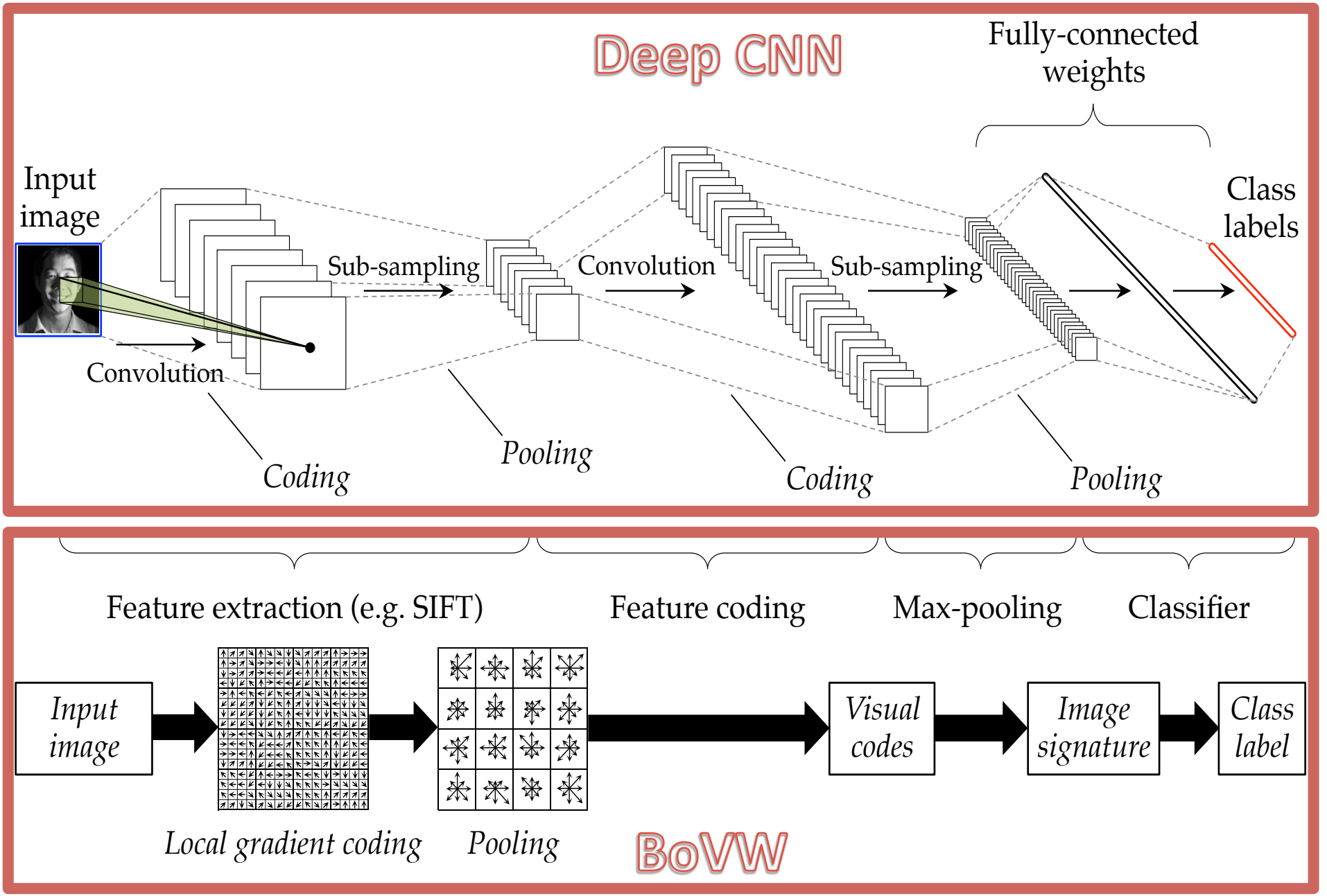



- Deep: learning hierarchical feature representations (disentangle multiple factors)
- Theo. Shallow model same potential as deep but:
  - Need (exponentially) more param
  - CNN Not a simple MLP => a lot of structures (knowledge) embedded via Convolutional layers
- Is it possible to Mimic a deep with a shallow architecture?

  [Do Deep Nets Really Need to be Deep?, Ba NIPS 2014]



# Feature engineering/ feature learning

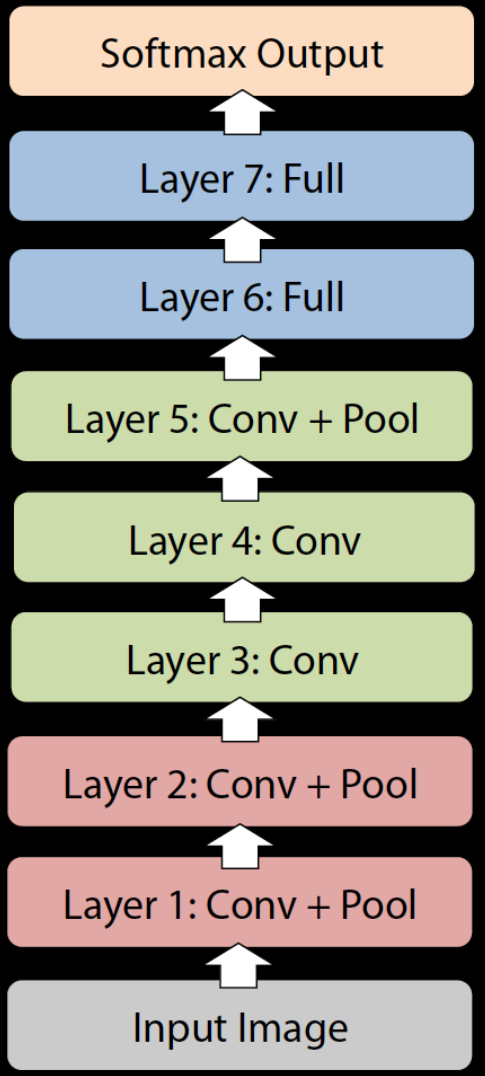


# Deep vs shallow in Computer Vision

- CV works a lot on handcrafted local features
  - BoVW (Bag of Visual Words and extensions FisherVectors, BossaNova …)
  - BoVW not so shallow but
    - not end-to-end supervised learning
- Deep CNN: end-to-end learning on a handcrafted architecture! [Chatfield BMVC 2014]
  - Why 8 layers? why 3x3 at the 5th layer without polling? … => ad-hoc architecture

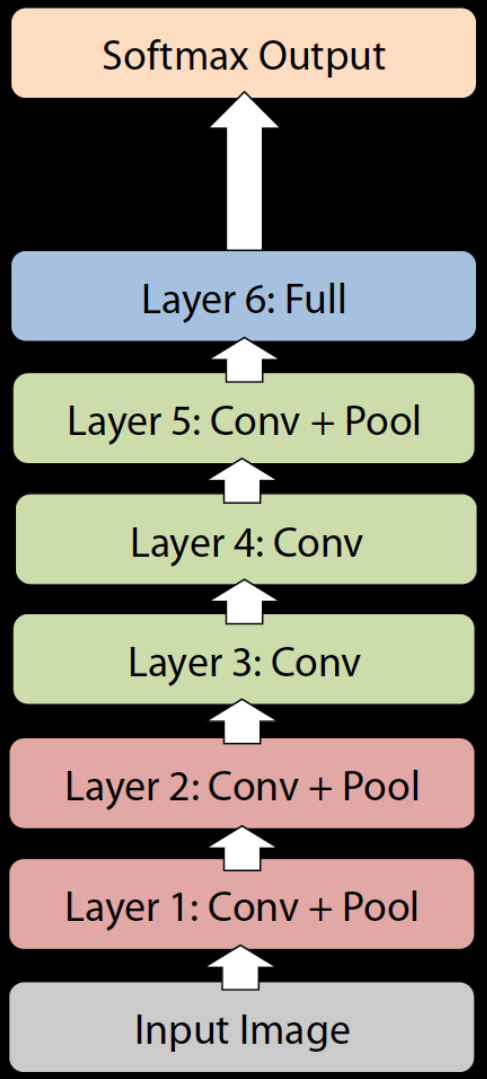
# Architecture of Krizhevsky et al.

- 8 layers total
- Trained on Imagenet dataset [Deng et al. CVPR'09]
- 18.2% top-5 error
- Our reimplementation:
  18.1% top-5 error



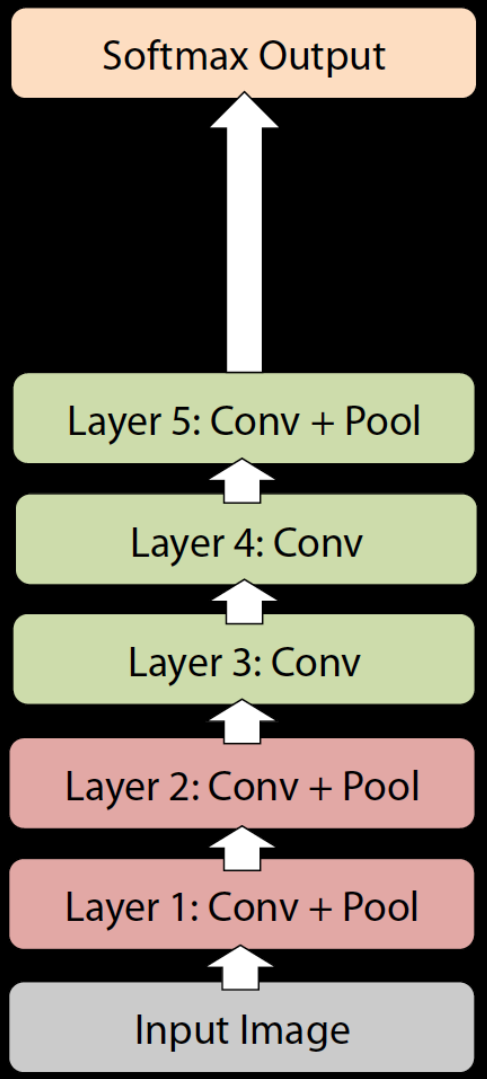
# Architecture of Krizhevsky et al.

- Remove top fully connected layer
  - Layer 7
- Drop 16 million parameters
- Only 1.1% drop in performance!



# Architecture of Krizhevsky et al.

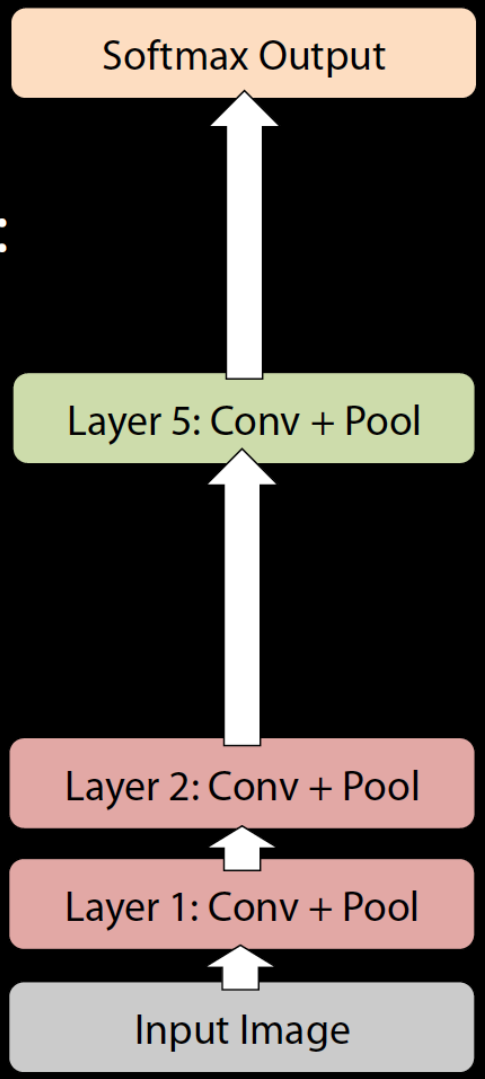
- Remove both fully connected layers
  - Layer 6 & 7
- Drop ~50 million parameters
- 5.7% drop in performance



# Architecture of Krizhevsky et al.

- Now try removing upper feature extractor layers & fully connected:
  - Layers 3, 4, 6 ,7
- Now only 4 layers
- 33.5% drop in performance

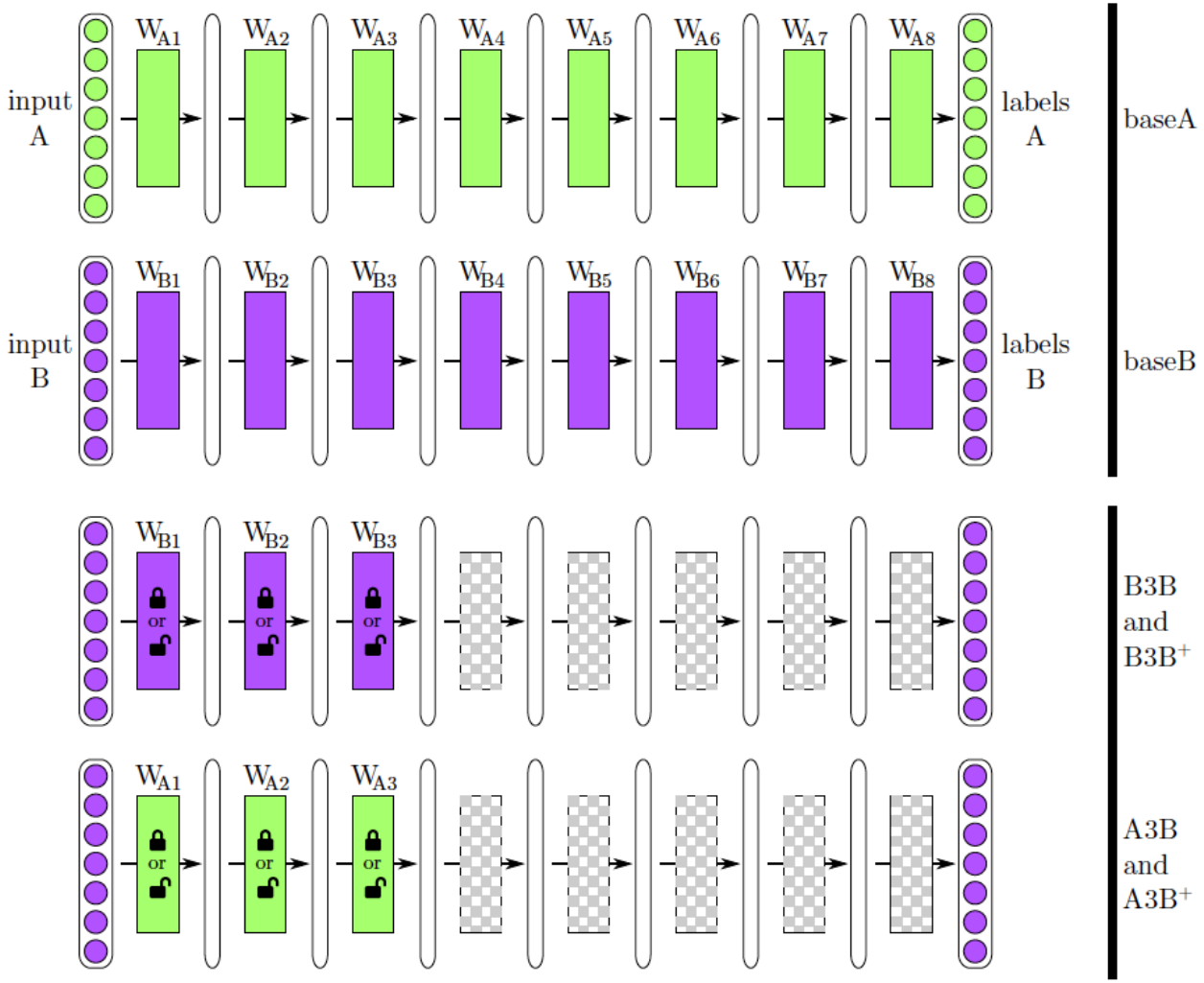
→Depth of network is key



# Using CNN representation in Vision

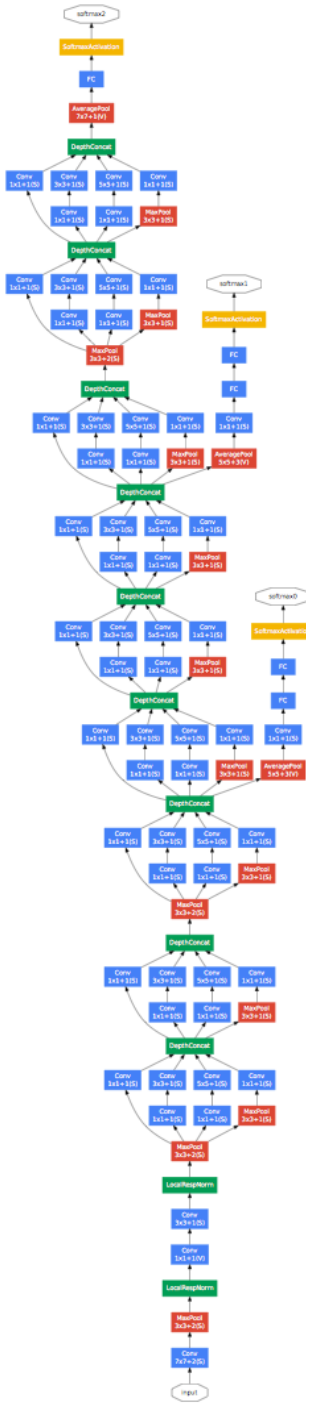
- Are CNN providing generic features ?
  - Yes! Deep features (from ImageNet) +SVM on PASCAL 07 == 10% better than best BoVW methods! [Chatfield]
- Transfert to many tasks [Razavian CVPRw2014]
  - Frozen features + SVM = solution to small datasets
  - Fine tuning not easy in that case (small datasets)
  - Which is the best layer cut for transfert ?
    - Depending to the task

# How it is Transferable? [Yosinski NIPS 2014]



# To conclude (1/2)

GoogLeNet
20 layers
Supervision at multiple layers
Err = 6.6%



- Supervised/Unsupervised – learning generic data representation
- Incorporate domain knowledge into deep architectures
- Weak on theoretical support:
  - convergence bound,
  - local minimum,
  - why it works ???

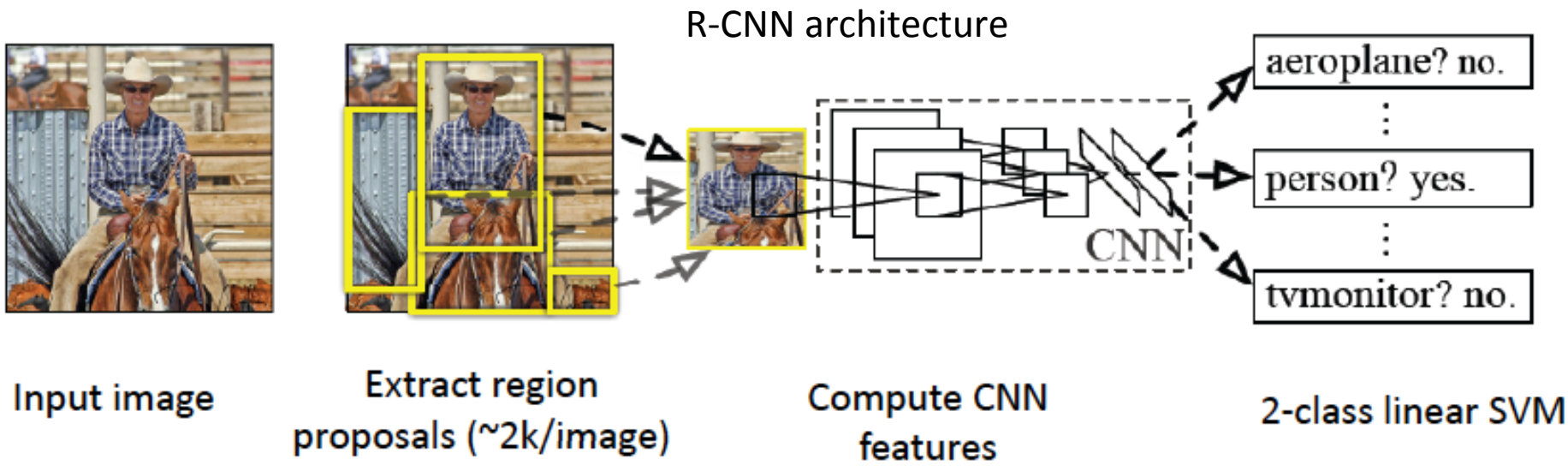
  ⇒deep structure analysis/understanding

  ⇒Next Talk of S. Mallat --Filter banks / Scattering / contractive architecture …

# To conclude (2/2)

- Computer Vision side:
  - ImageNet: Object recognition task
    - How to do for large and complex scenes ?
    - R-CNN [Girshik CVPR2014]

⇒ to be explored in the talk of I. Laptev



# LIP6 Team Ref. on deep learning and Visual representation

Deep learning for Visual Recognition

- Top-Down Regularization of Deep Belief Networks, H. Goh, N. Thome, M. Cord, JH. Lim, NIPS 2013
- Learning Deep Hierarchical Visual Feature Coding, H. Goh+, IEEE Transactions on Neural Networks and Learning Systems 2014
- Unsupervised and supervised visual codes with Restricted Boltzmann Machines, H. Goh+, ECCV 2012
- Learning invariant color features with sparse topographic restricted boltzmann machines, H. Goh+, ICIP 2011
- Biasing Restricted Boltzmann Machines to Manipulate Latent Selectivity and Sparsity, H. Goh+, NIPS workshop 2010

Bio-inspired Representation

- Extended coding and pooling in the HMAX model, C. Thériault, N. Thome, M. Cord, IEEE Trans. on Image Processing 2013

Visual representation

- Pooling in Image Representation: the Visual Codeword Point of View, S. Avila, N. Thome, M. Cord, E. Valle, A. araujo, CVIU 2013
- Dynamic Scene Classification: Learning Motion Descriptors with Slow Features Analysis, C. Thériault, N. Thome, M. Cord, CVPR 2013

Matthieu Cord
LIP6, Computer Science Department
UPMC Paris 6 - Sorbonne University
Paris FRANCE
http://webia.lip6.fr/~cord