Deep learning and weak supervision for image classification

Matthieu Cord (LIP6 UPMC)
Joint work with Thibaut Durand and Nicolas Thome
MLIA Team (Patrick Gallinari)

Université Pierre et Marie Curie (UPMC / Paris 6 / Jussieu) - LIP6
Sorbonne Universités

17th of May 2016

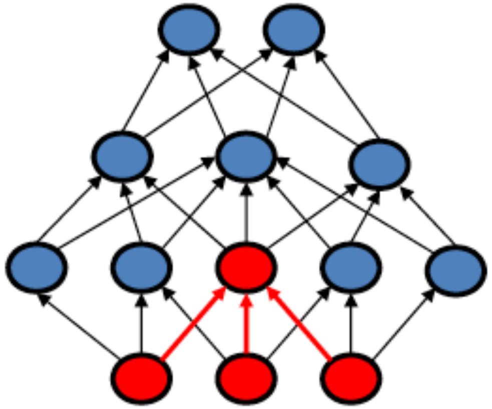
Outline

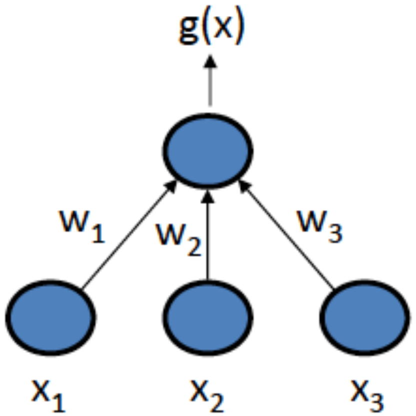
1. Deep learning for object recognition
   - Architecture
   - Results
   - Learning
   - Using deep in Computer Vision
   - Key issues for Deep & Vision
2. Weakly supervised deep learning
   - Architecture
   - Results

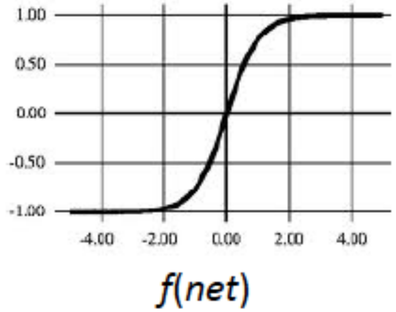
Neural network
Back propagation

Nature

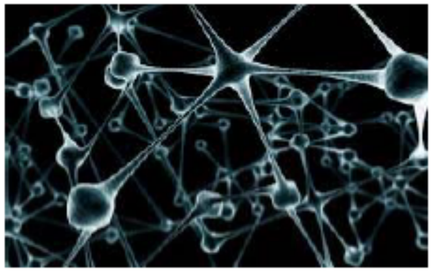
1986







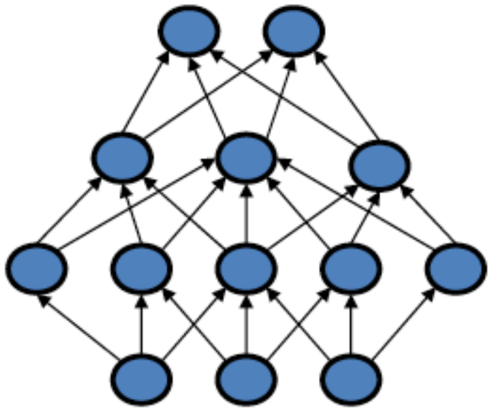
$$g(\mathbf{x}) = f(\sum_{i=1}^{d} x_i w_i + w_0) = f(\mathbf{w}^t \mathbf{x})$$



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

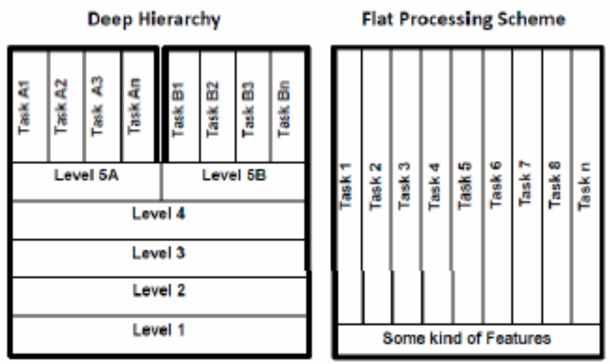
Neural network
Back propagation

Nature

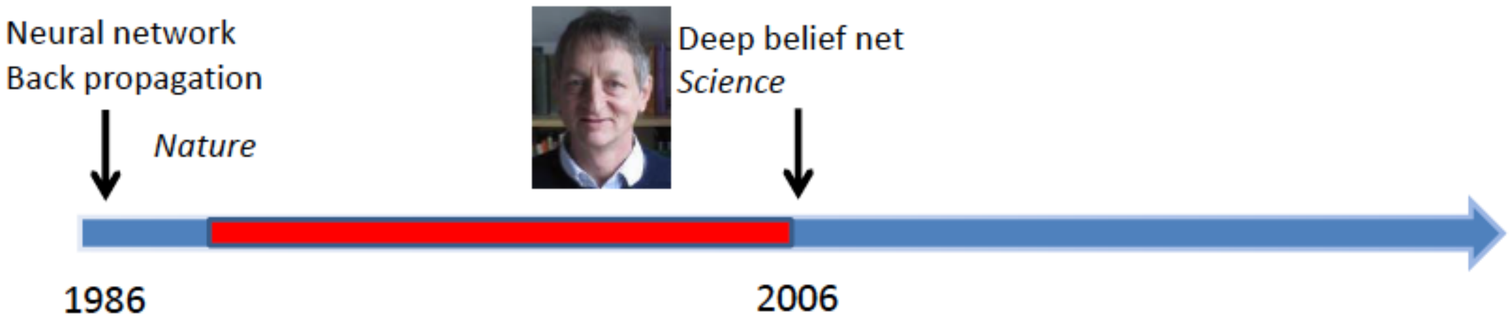
1986 Deep winter 2006

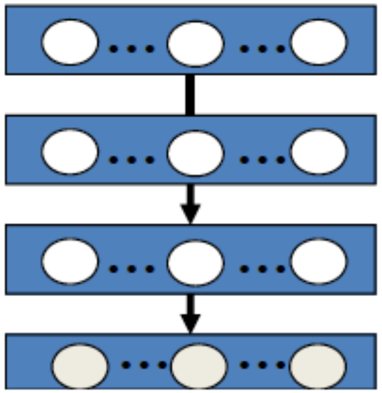
- SVM
- Boosting
- Decision tree
- KNN
- …

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



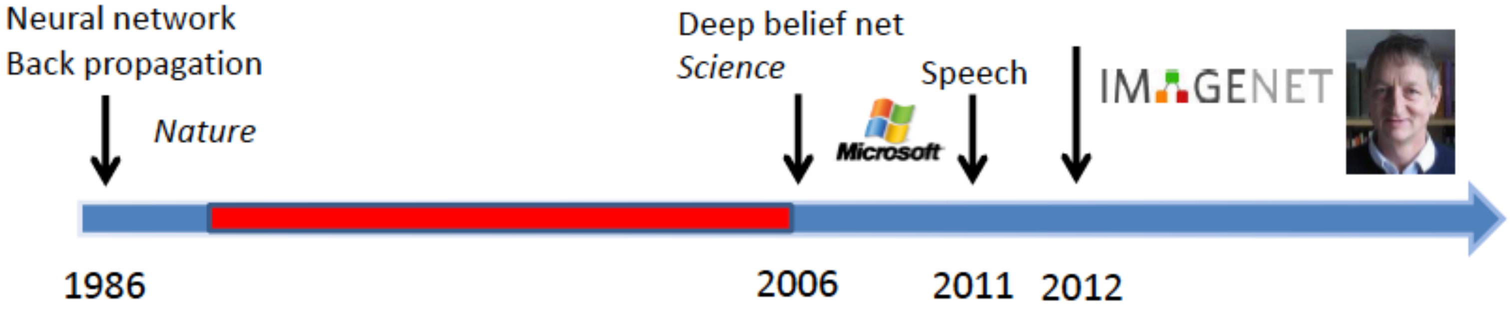
Kruger et al. TPAMI'13





- 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

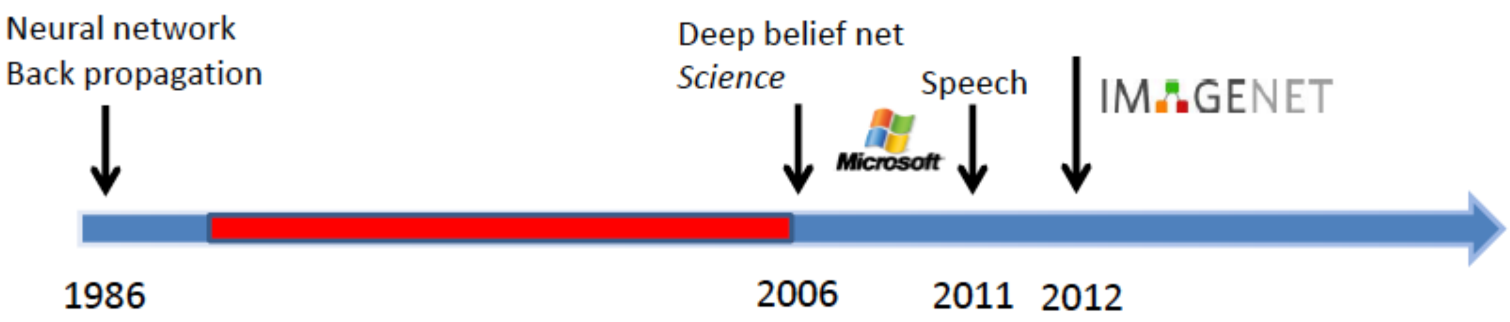
Big Data !



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.



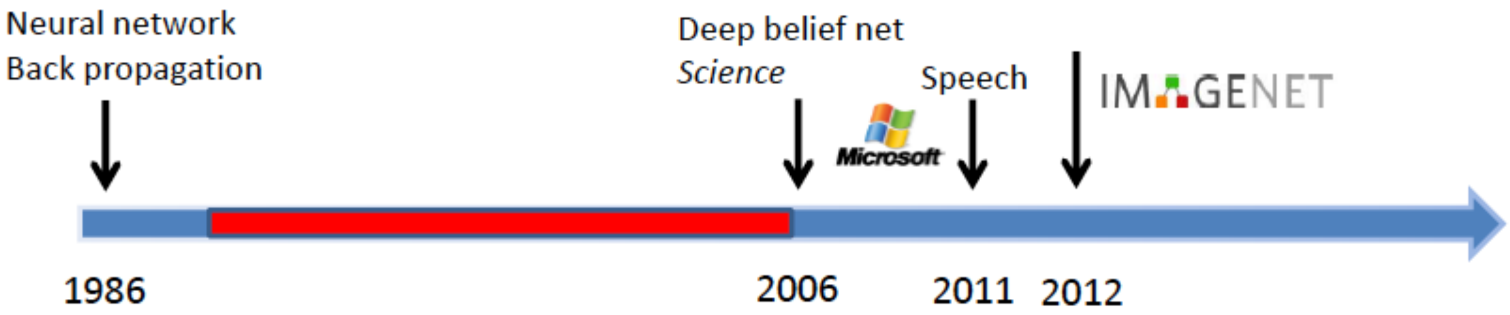
- ImageNet 2013 – image classification challenge

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

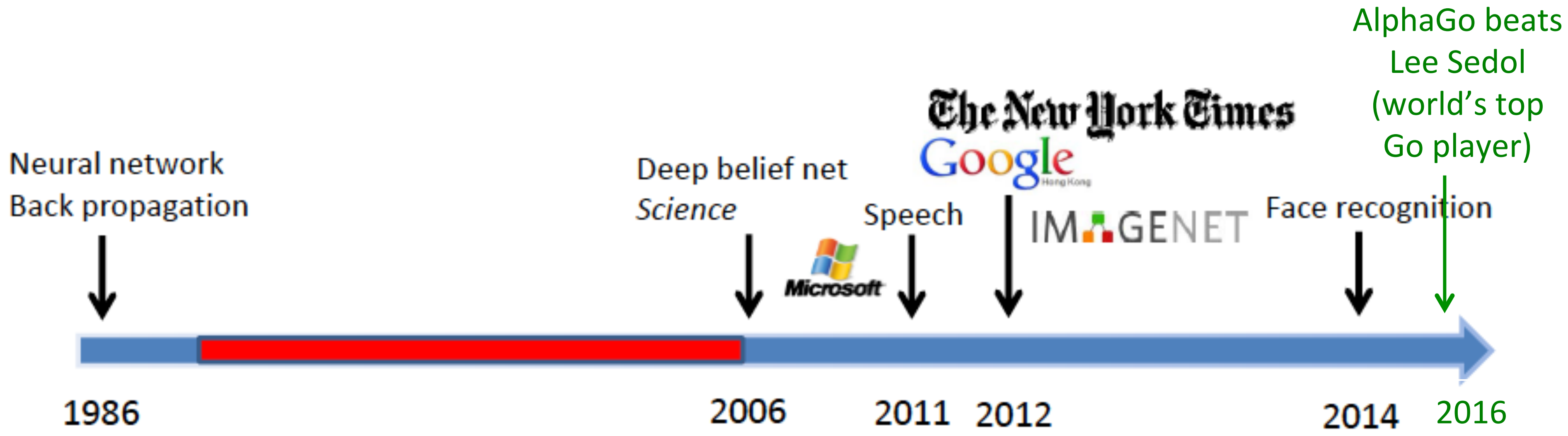


- ImageNet 2014 – Image classification challenge

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

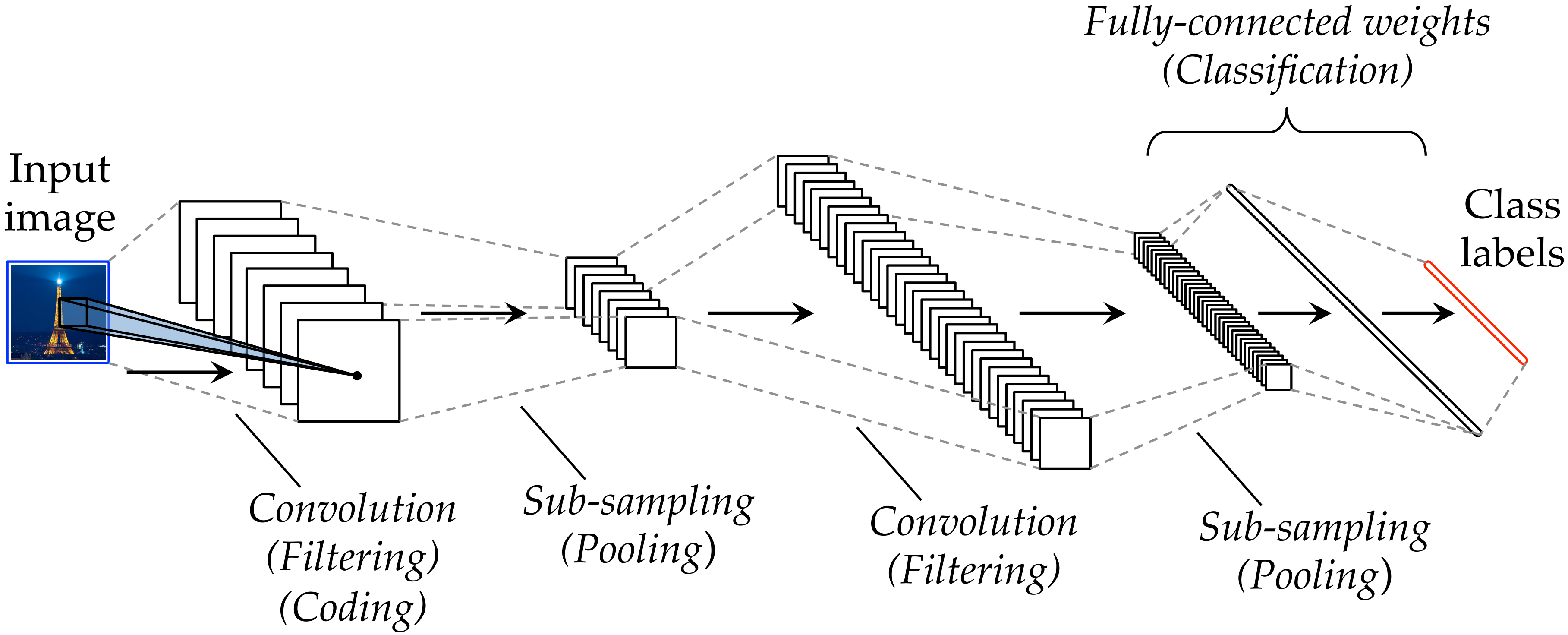
Rank	Name	Mean Average Precision	Description
1	Google	0.43933	Deep learning
2	CUHK	0.40656	Deep learning
3	DeepInsight	0.40452	Deep learning
4	UvA-Euvision	0.35421	Deep learning
5	Berkley Vision	0.34521	Deep learning



Neural network
Back propagation
Deep belief net
Science
Speech
Face recognition
AlphaGo beats
Lee Sedol
(world's top
Go player)
1986
2006
2011
2012
2014
2016

Deep Convolutional Neural Networks (CNN)

[LeCun-89]

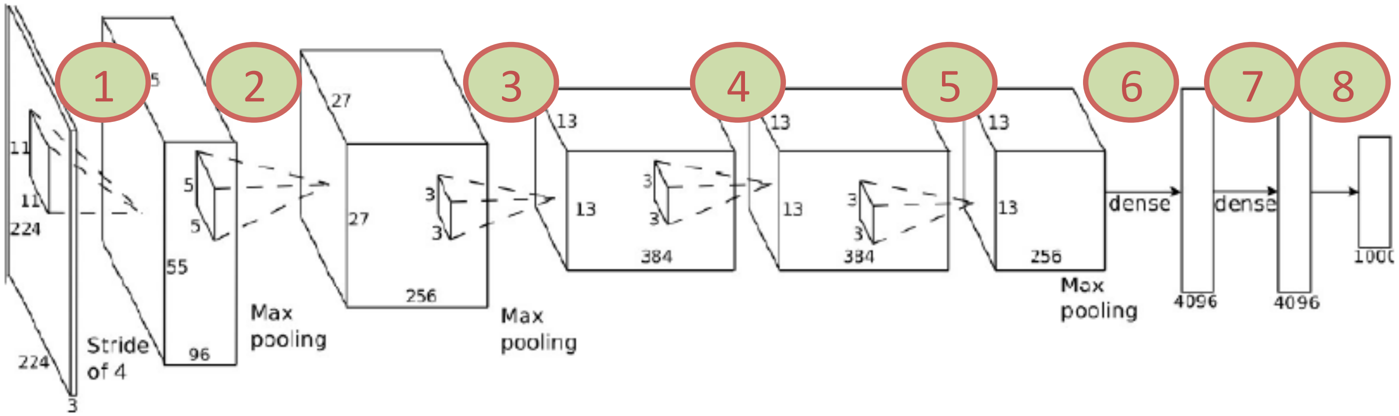


- **Convolution** uses local weights shared across the whole image
- **Pooling** shrinks the spatial dimensions
- Many other Deep Models (not convolutional):
  - Deep belief Net – Hinton'06 – Stack RBM
  - Auto-Encoder – Hinton and Salakhutdinov 06

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

Krizhevsky et al. [NIPS2012]

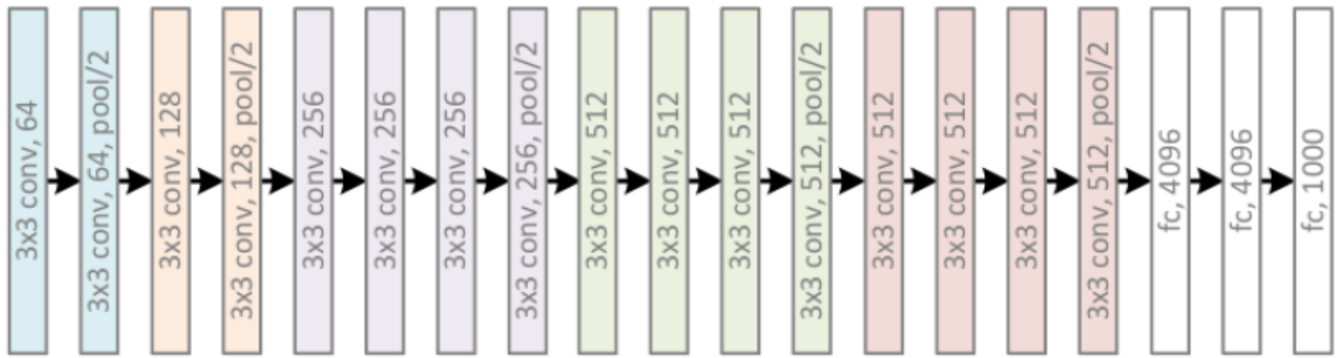
- Same model as LeCun'98 but:
  - Bigger model (8 layers)
  - More data (10^6 vs 10^3 images)
  - GPU implementation (50x speedup over CPU)
  - Better regularization (DropOut)



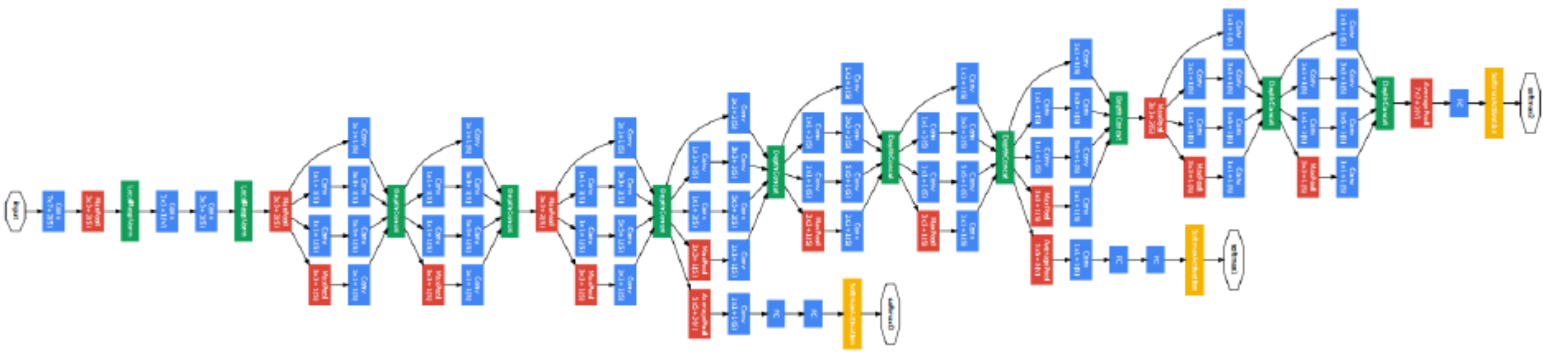
- 7 hidden layers, 650,000 neurons, 60,000,000 parameters
- Trained on 2 GPUs for a week

From very deep to very very very …

VGG, 16/19 layers, 2014

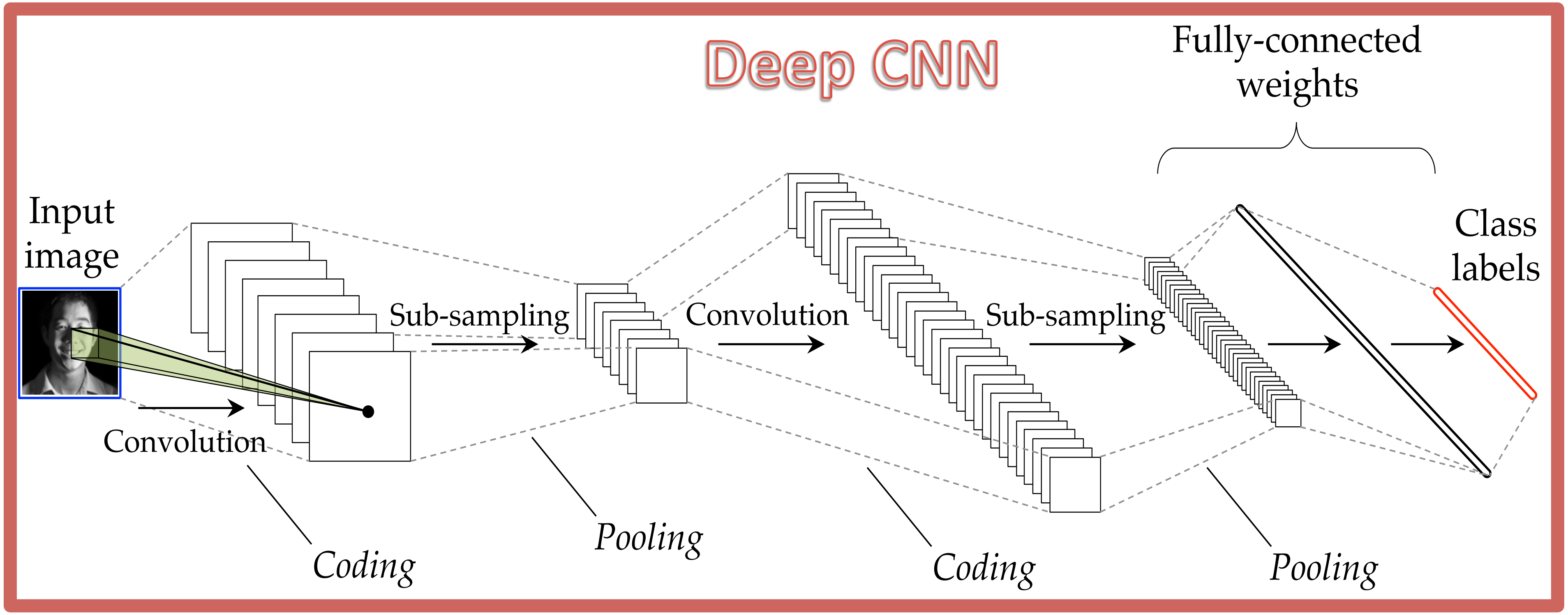


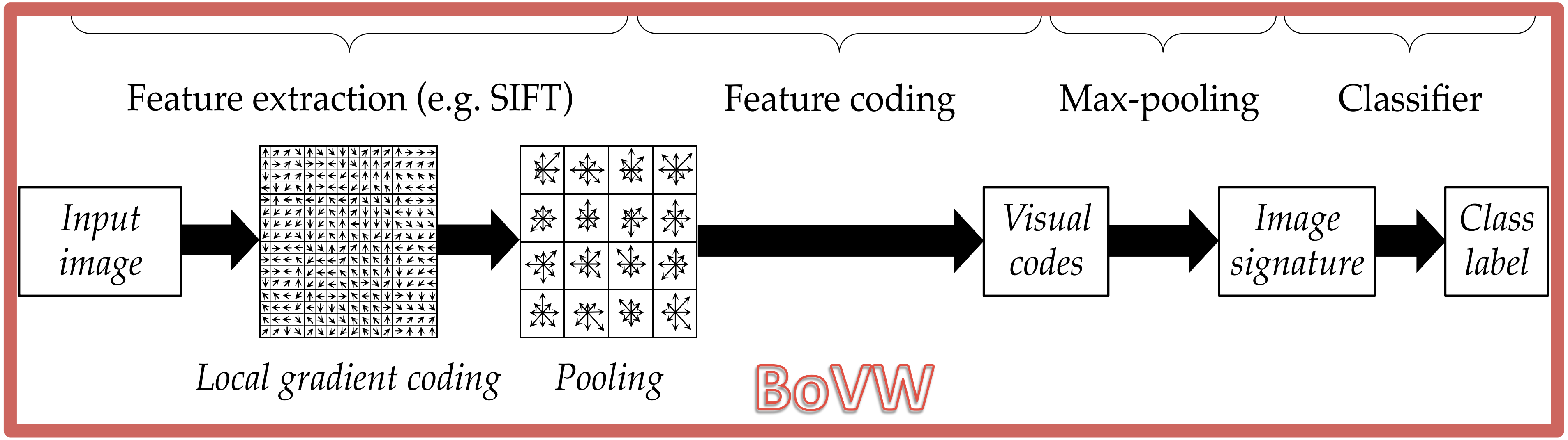
GoogleNet, 22 layers, 2014



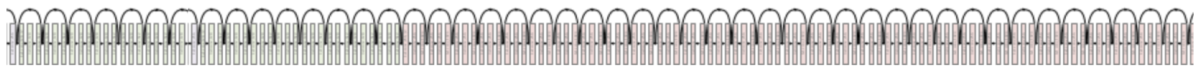
ResNet, 152 layers, 2015

Feature engineering/ feature learning





Key issues for Deep&Vision

- Supervised/Unsupervised(predictive) – learning generic data representation
  - ⇒ “*L'apprentissage profond non-supervisé : questions ouvertes*”, Y. LeCun (Facebook, NYU), 2nd workshop on deep (2016) at LIP6 with GdR ISIS
- Weak on theoretical support:
  - Convergence => math of deep learning tuto Vidal/Bruna ICCV 2015
  - Why it works ? Deep structure analysis/understanding
  - ⇒ Talk of S. Mallat (Collège de France 2016): “on y comprend à peu près rien”, first workshop on deep (2015) at LIP6 with GdR ISIS
  - How many layers ? => 
- ImageNet: Object recognition task
  - How to do for large and complex scenes ?
  - Localization: R-CNN [Girshick CVPR2014]
  - ⇒ Fast R-CNN [ICCV 2015], Faster R-CNN [NIPS 2015]

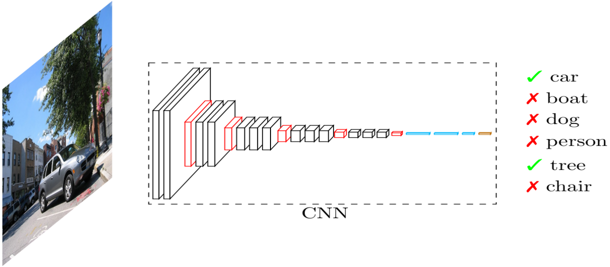
Outline

1. Deep learning for object recognition
   - Architecture
   - Results
   - Learning
   - Using deep in Computer Vision
   - Key issues for Deep & Vision
2. **WSL deep learning**

Joint work with Thibaut Durand and Nicolas Thome (LIP6)

Context: image classification

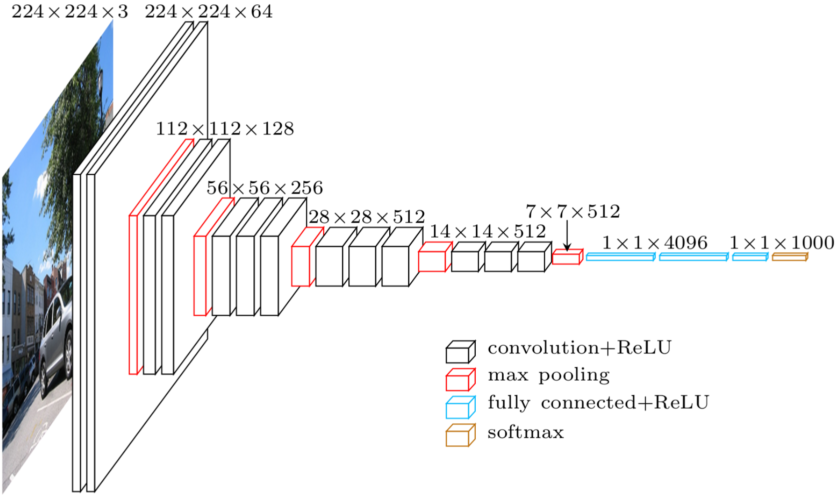
- Xlabels
- Deep CNN



- Weakly-Supervised Learning (WSL)
- Select relevant regions → better prediction
- No bounding box (expensive)
- Baseline model: Latent SVM [Felzenszwalb, PAMI10]

label="car"

Standard deep CNN architecture: VGG16

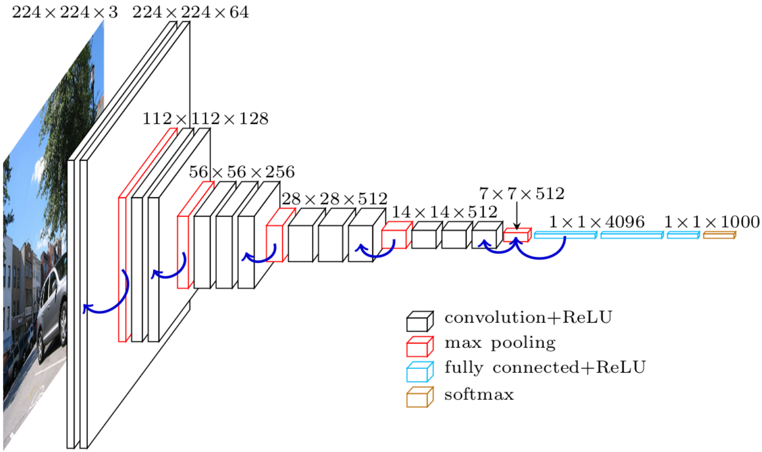


Simonyan et al. *Very deep convolutional networks for large-scale image recognition.* ICLR 2015

WSL adaptation for deep CNN

Problem

- Fixed-size image as input



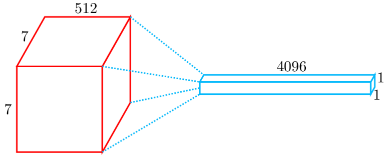
WSL adaptation for deep CNN

Problem

- Fixed-size image as input

Adapt architecture to weakly supervised learning

1. Fully connected layers → convolution layers
   - sliding window approach



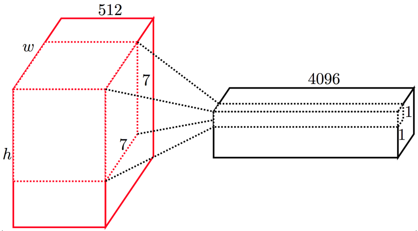
WSL adaptation for deep CNN

Problem

- Fixed-size image as input

Adapt architecture to weakly supervised learning

1. Fully connected layers $\rightarrow$ convolution layers
   - sliding window approach



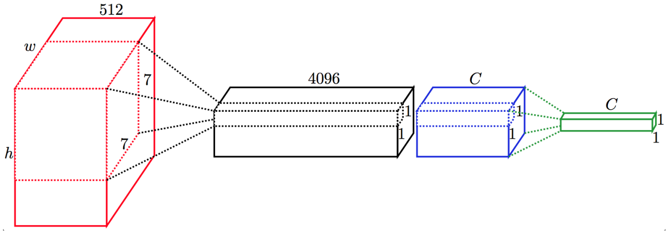
WSL adaptation for deep CNN

Problem

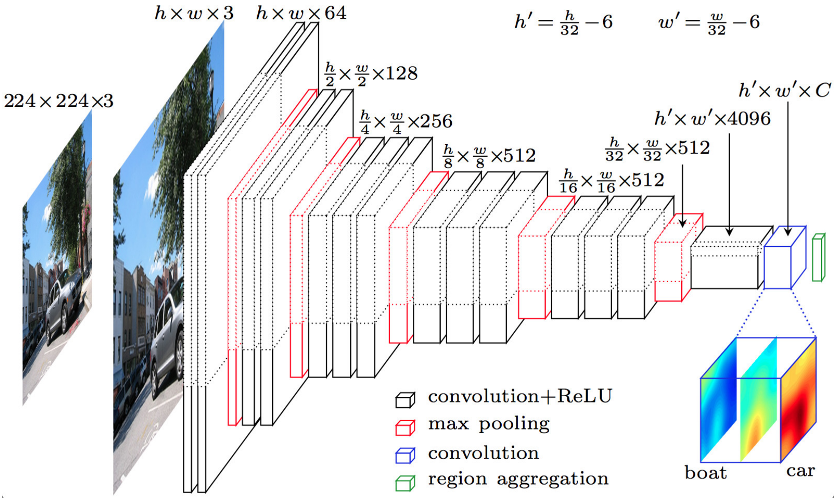
- Fixed-size image as input

Adapt architecture to weakly supervised learning

1. Fully connected layers $\rightarrow$ convolution layers
   - sliding window approach
2. Spatial aggregation
   - Perform object localization prediction

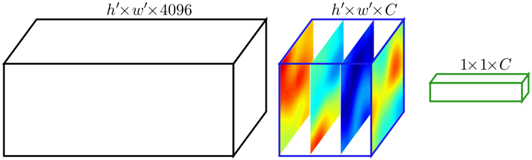


WSL deep architecture



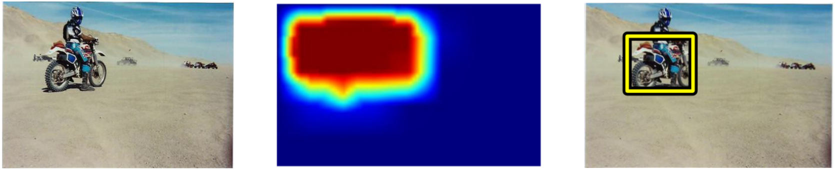
- C: number of classes

WSL deep architecture: region selection



[Oquab, 2015]

- Region selection = max
- Select the highest-scoring window



original image | **motorbike** feature map | max prediction

Oquab, Bottou, Laptev, Sivic. *Is object localization for free? weakly-supervised learning with convolutional neural networks.* CVPR 2015

Our WSL deep CNN: region selection

New region selection strategy

- max + min pooling (MANTRA prediction function)
  - max: indicator of the **presence** of the class
  - min: indicator of the **absence** of the class
- Use negative evidence

street image | **street** model | **coast** model | **highway** model

- Durand et al. *MANTRA: minimum maximum latent structural svm.* ICCV 2015
- Parizi et al. *Automatic discovery and optimization of parts.* ICLR 2015

Our WSL deep CNN: region selection

k-instances

- Single region to multiple high scoring regions:

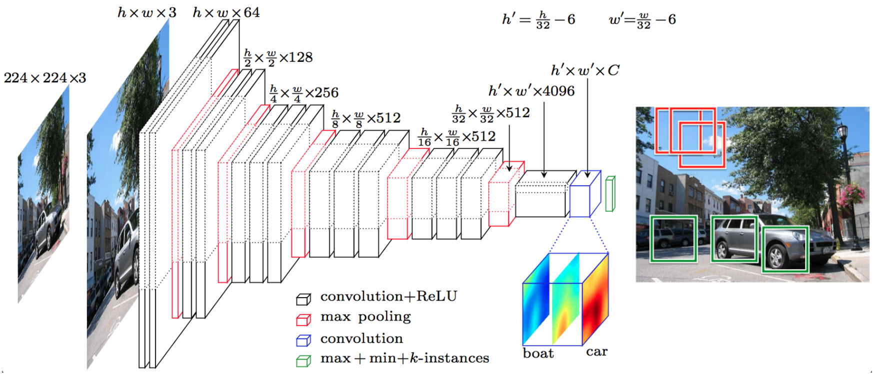
$$\max \rightarrow \frac{1}{k}\sum_{i=1}^{k} i\text{-th}\max$$

- More robust region selection.


k=1


k=3

Our WSL deep CNN: architecture

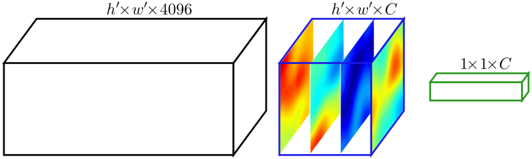


Our WSL deep CNN: learning

- Objective function for multi-class task and $k = 1$:

$$\min_{\mathbf{w}} \mathcal{R}(\mathbf{w}) + \frac{1}{N} \sum_{i=1}^{N} \ell(f_{\mathbf{w}}(\mathbf{x}_i), y_i^{gt})$$

$$f_{\mathbf{w}}(\mathbf{x}_i) = \arg\max_{y} \left(\max_{h} \mathbf{L}_{\mathbf{conv}}^{\mathbf{w}}(\mathbf{x}_i, y, h) + \min_{h'} \mathbf{L}_{\mathbf{conv}}^{\mathbf{w}}(\mathbf{x}_i, y, h') \right)$$



How to learn deep architecture ?

- Stochastic gradient descent training.
- Back-propagation of the **selecting windows** error.

Our WSL deep CNN: learning

Class is **present**

- **Increase** score of selecting windows.

Figure: Car map

Our WSL deep CNN: learning

Class is **absent**

- **Decrease** score of selecting windows.

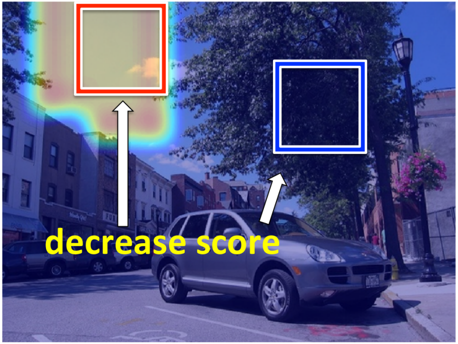


Figure: Boat map

Experiments

- VGG16 pre-trained on ImageNet
- Torch7 implementation

Datasets

- Object recognition: Pascal VOC 2007, Pascal VOC 2012
- Scene recognition: MIT67, 15 Scene
- Visual recognition, where context plays an important role: COCO, Pascal VOC 2012 Action

VOC07/12

MIT67

15 Scene

COCO

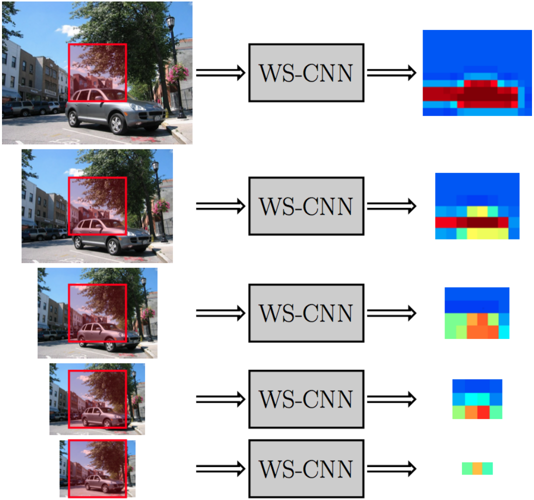
VOC12 Action

Experiments

Dataset	Train	Test	Classes	Classification
VOC07	$\sim$5.000	$\sim$5.000	20	multi-label
VOC12	$\sim$5.700	$\sim$5.800	20	multi-label
15 Scene	1.500	2.985	15	multi-class
MIT67	5.360	1.340	67	multi-class
VOC12 Action	$\sim$2.000	$\sim$2.000	10	multi-label
COCO	$\sim$80.000	$\sim$40.000	80	multi-label

Experiments

- Multi-scale: 8 scales (combination with Object Bank strategy)



Object recognition

	VOC 2007	VOC 2012
VGG16 (online code) [1]	84.5	82.8
SPP net [2]	82.4	
Deep WSL MIL [3]		81.8
Our WSL deep CNN	**90.2**	**88.5**

Table: mAP results on object recognition datasets.

[1] Simonyan et al. *Very deep convolutional networks.* ICLR 2015
[2] He et al. *Spatial pyramid pooling in deep convolutional networks.* ECCV 2014
[3] Oquab et al. *Is object localization for free?* CVPR 2015

Scene recognition

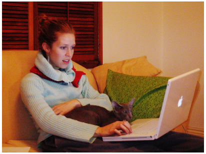
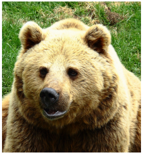
    

	15 Scene	MIT67
VGG16 (online code) [1]	91.2	69.9
MOP CNN [2]		68.9
Negative parts [3]		77.1
Our WSL deep CNN	**94.3**	**78.0**

Table: Multi-class accuracy results on scene categorization datasets.

[1] Simonyan et al. *Very deep convolutional networks.* ICLR 2015
[2] Gong et al. *Multi-scale Orderless Pooling of Deep Convolutional Activation Features.* ECCV 2014
[3] Parizi et al. *Automatic discovery and optimization of parts.* ICLR 2015

Context datasets

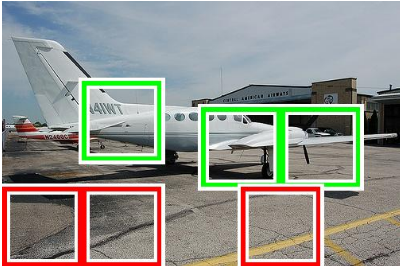
   

	VOC 2012 action	COCO
VGG16 (online code) [1]	67.1	59.7
Deep WSL MIL [2]		62.8
Our WSL deep CNN	**75.0**	**68.8**

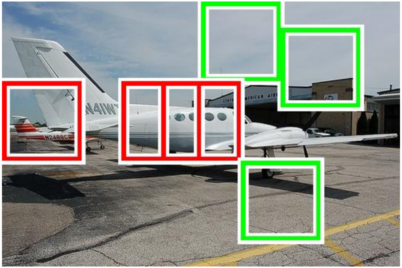
Table: mAP results on context datasets.

[1] Simonyan et al. *Very deep convolutional networks.* ICLR 2015
[2] Oquab et al. *Is object localization for free?* CVPR 2015

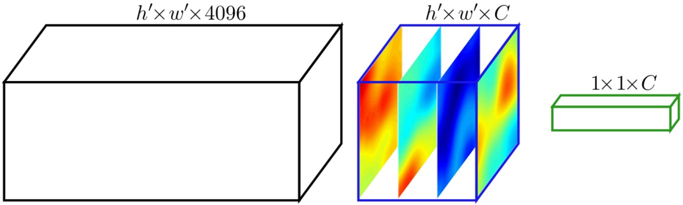
Visual results



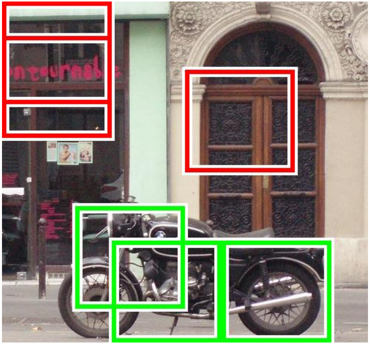
Aeroplane model (1.8)



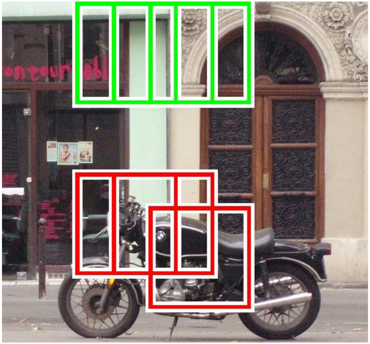
Bus model (-0.4)



Visual results



Motorbike model (1.1)



Sofa model (-0.8)

Visual results

Sofa model (1.2)

Horse model (-0.6)

Analysis

- Analyze the different improvements.
- Mono-scale experiments (smallest).

a) `max`	b) +k=3	c) +`min`	d) +AP	VOC07	VOC12 action
✓				83.6	53.5
✓	✓			86.3	62.6
✓		✓		87.5	68.4
✓		✓	✓	88.4	71.7
✓	✓	✓		87.8	69.8
✓	✓	✓	✓	**88.9**	**72.6**

Analysis

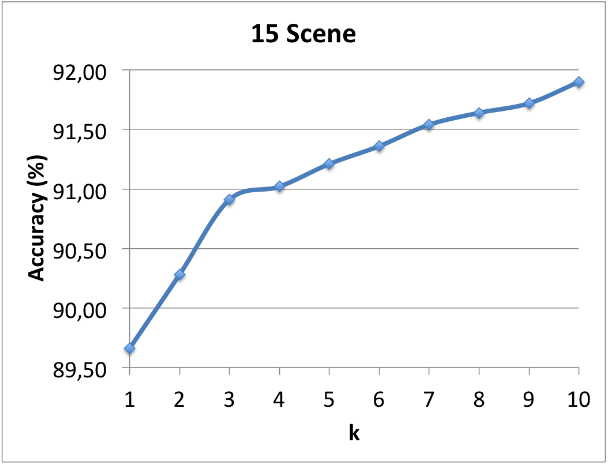
- Analyze the different improvements.
- Mono-scale experiments (smallest).

a) `max`	b) +k=3	c) +`min`	d) +AP	VOC07	VOC12 action
✓				83.6	53.5
✓	✓			86.3	62.6
✓		✓		87.5	68.4
✓		✓	✓	88.4	71.7
✓	✓	✓		87.8	69.8
✓	✓	✓	✓	**88.9**	**72.6**

- max + min > max
- with top > without top
- AP loss > Acc loss

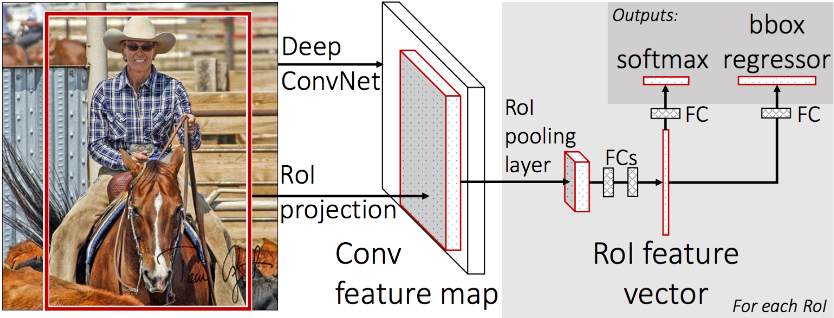
Analysis

- Impact of the number or regions k



Comparison with supervised object detector

- Fast Region-based Convolutional Network (Fast R-CNN)



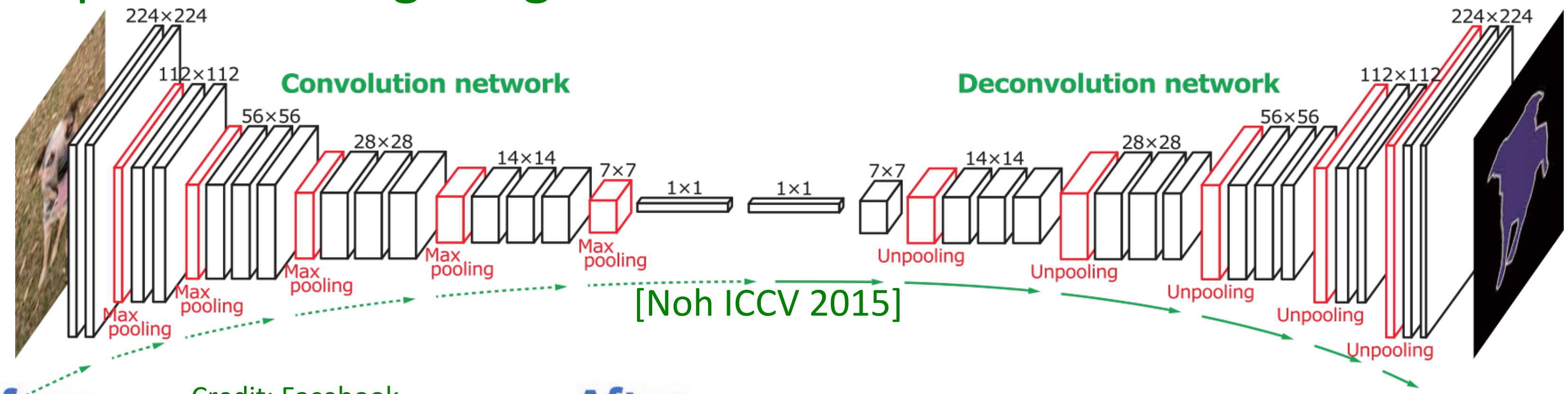
Girshick. *Fast R-CNN.* ICCV 2015

Comparison with supervised object detector

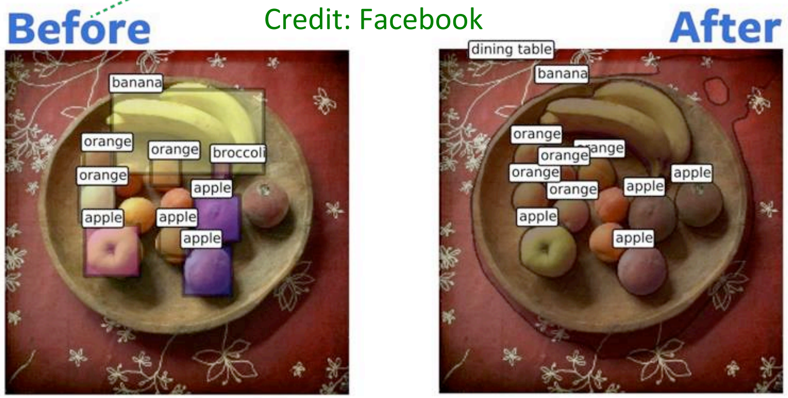
	Fast R-CNN	WS-CNN
Goal	detection	classification
Region proposal	selective search image dependent several ratio/size	sliding window fixed grid 1 ratio/size per scale
Region pooling	RoI pooling layer	network architecture
Forward	all regions	all regions
Back-propagation	all regions	only selected regions
Loss	region-level	image-level
Annotation	object bounding-boxes	presence/absence

(more) Key issues for Deep&Vision

- Supervised Image Segmentation task



[Noh ICCV 2015]



Credit: Facebook

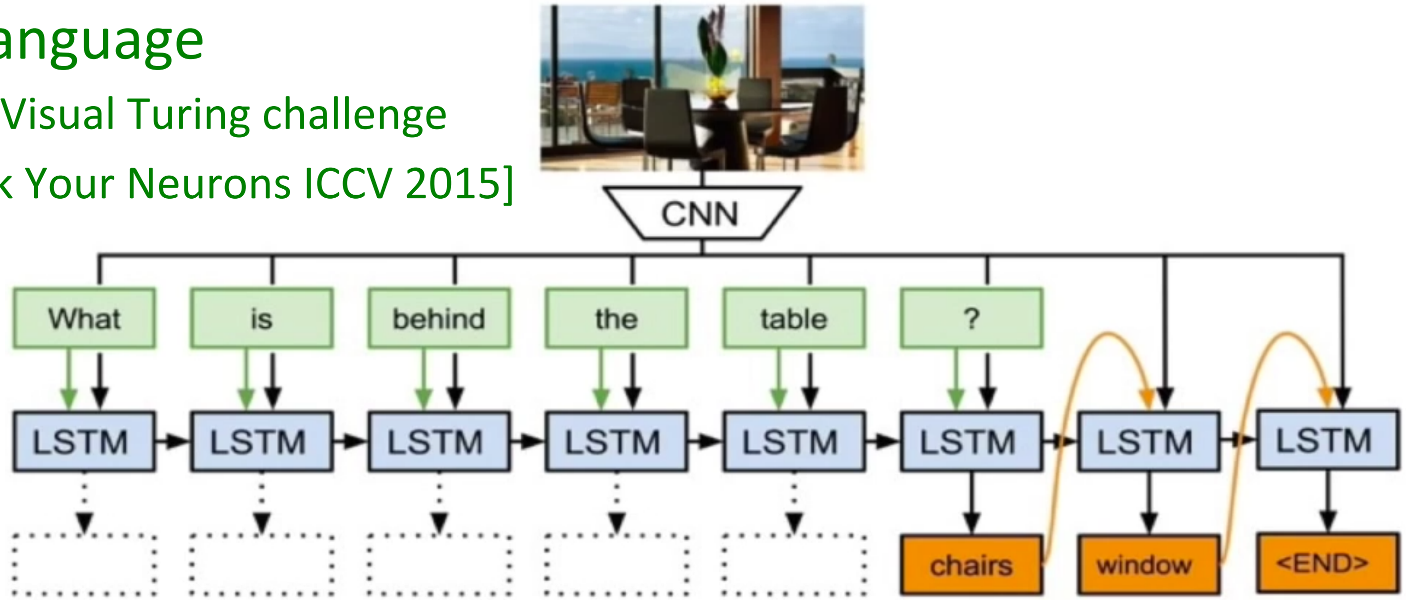
MS COCO Detection Challenge!

- Deep generative models
- Compression/Embedded/Green nets

(more) Key issues for Deep&Vision

- Vision and Language
  - Visual Q&A, Visual Turing challenge

  [Malinowski Ask Your Neurons ICCV 2015]



  - Visual7W: Grounded Question Answering in Images [Yuke Zhu…Fei-Fei CVPR 16]

- Connection to sequential learning RNN, LSTM, memory nets, …
- Connection to Neurosciences

LIP6 Team Ref. on deep learning and Visual representation:

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

Deep learning for Visual Recognition

- WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks, T. Durand, N. Thome, M. Cord, CVPR 2016
- Deep Neural Networks Under Stress, M. Carvalho, M. Cord, S. Avila, N. Thome, E. Valle, ICIP 2016
- Max-Min convolutional neural networks for image classification, M. Blot, M. Cord, N. Thome, ICIP 2016
- MANTRA: Minimum Maximum Latent Structural SVM for Image Classification and Ranking, T Durand, N Thome, M Cord, ICCV 2015
- LR-CNN for fine-grained classification with varying resolution, M Chevalier, N Thome, M Cord, J Fournier, G Henaff, E Dusch, ICIP 2015
- Top-Down Regularization of Deep Belief Networks, H. Goh, N. Thome, M. Cord, JH. Lim, NIPS 2013
- Sequentially generated instance-dependent image representations for classification, G Dulac-Arnold, L Denoyer, N Thome, M Cord, P Gallinari, ICLR 2014
- 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
- Biasing Restricted Boltzmann Machines to Manipulate Latent Selectivity and Sparsity, H. Goh+, NIPS workshop 2010

Bio-inspired Representation

- Cortical Networks of Visual RecognitionC Thériault, N Thome, M Cord, Biologically Inspired Computer Vision: Fundamentals and Applications, book chapter
- 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