Apprentissage, réseaux de neurones et modèles graphiques (RCP209)
Neural Networks and Deep Learning

Nicolas Thome
Prenom.Nom@cnam.fr
`http://cedric.cnam.fr/vertigo/Cours/ml2/`

Département Informatique
Conservatoire Nationnal des Arts et Métiers (Cnam)

Outline

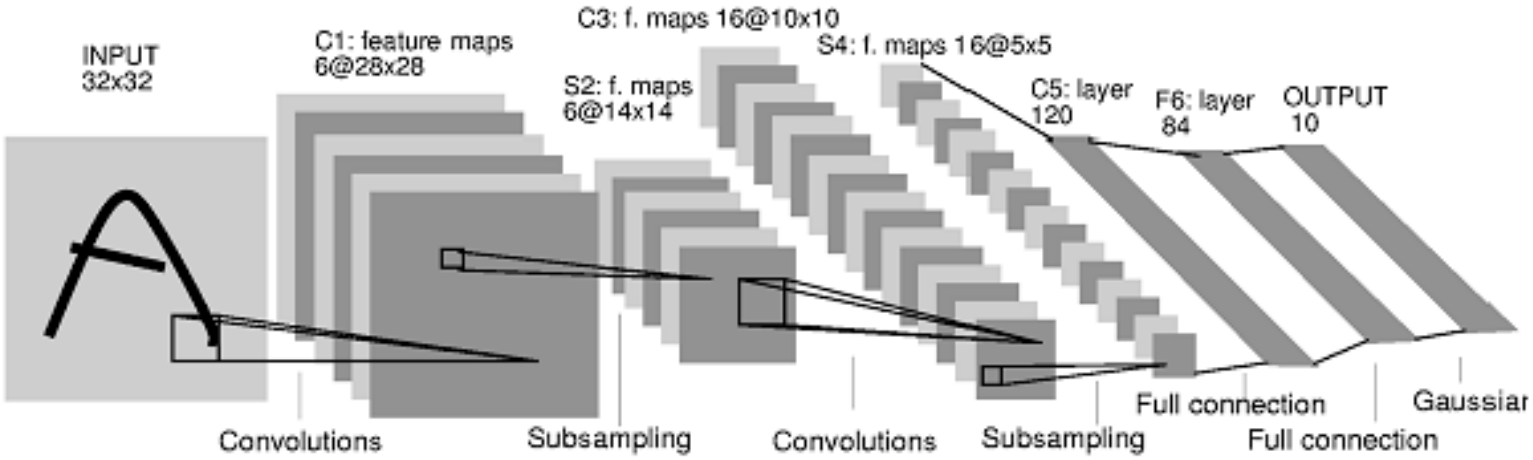
1 Deep Learning History

2 2012: the Deep Learning Revolution

Deep Learning: Trends and methods in the last four decades

80's: 1st Convolutionnal Neural Networks

- LeNet 5 Model [LBD^{+}89], trained using back-prop

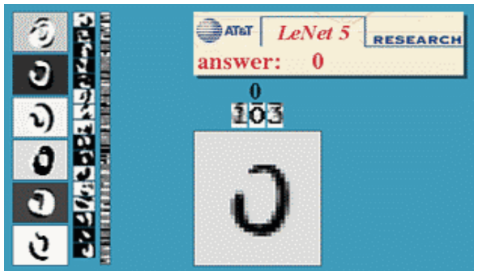


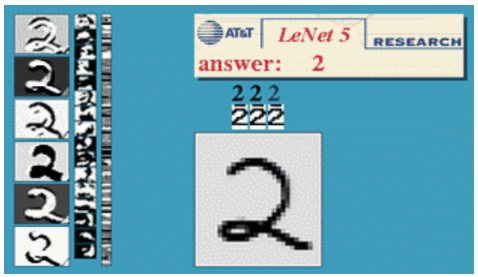
- Input: 32x32 pixel image. Largest character is 20x20
- 2 successive blocks [Convolution + Sigmoid + Pooling (+sigmoid)]
  Cx: Convolutional layer, Sx: Subsampling layer
- C5: convolution layer ~ fully connected
- 2 Fully connected layers Fx

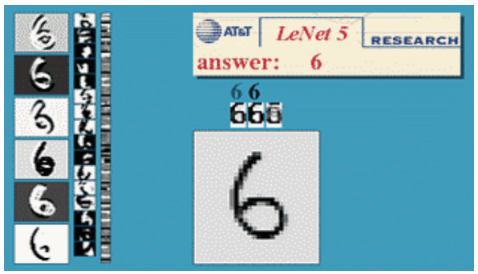
80's: LeNet 5 Model

- Evaluation on MNIST
- Total # parameters ~ 60000
  - 60,000 original datasets: test error: 0.95%
  - 540,000 artificial distortions + 60,000 original: Test error: 0.8%

- **Successful deployment for postal code reading in the US**



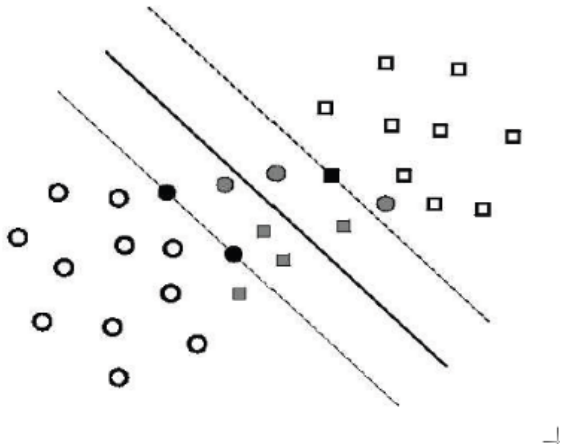


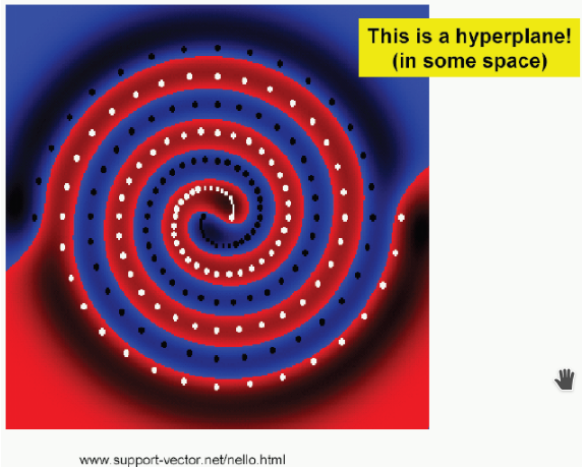


Deep Learning: Trends and methods in the last four decades

90's: start of winter for deep learning

- Deep neural nets =' black magic', black boxes
  - Lack of interpretability
  - Optimization issues for highly non-convex objective function
- **Golden age of kernel methods**
  - Generalization theory with Support Vector Machines
  - Extension to non-linear modes: kernel trick
    - Kernel encode prior knowledge (structure) on data
  - Convex optimization problem

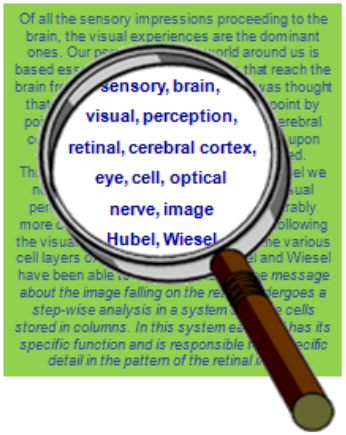


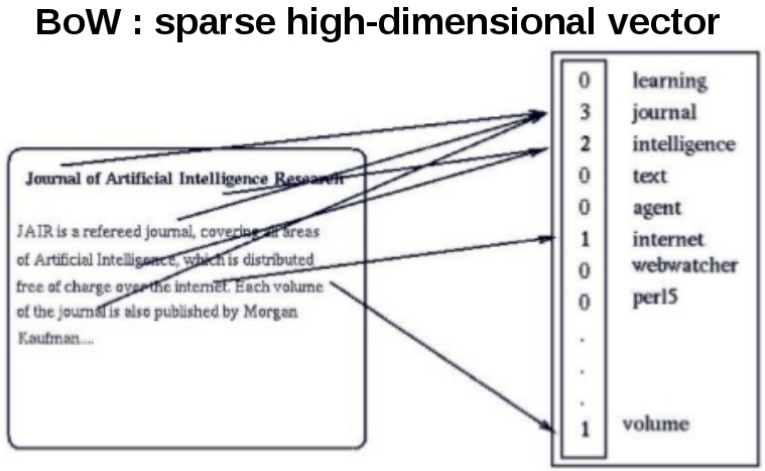


Deep Learning: Trends and methods in the last four decades

2000's: Bag of Words Model (BoW)

- Started from the Information Retrieval (IR) community
- Text classification : document as a histogram of word occurrences



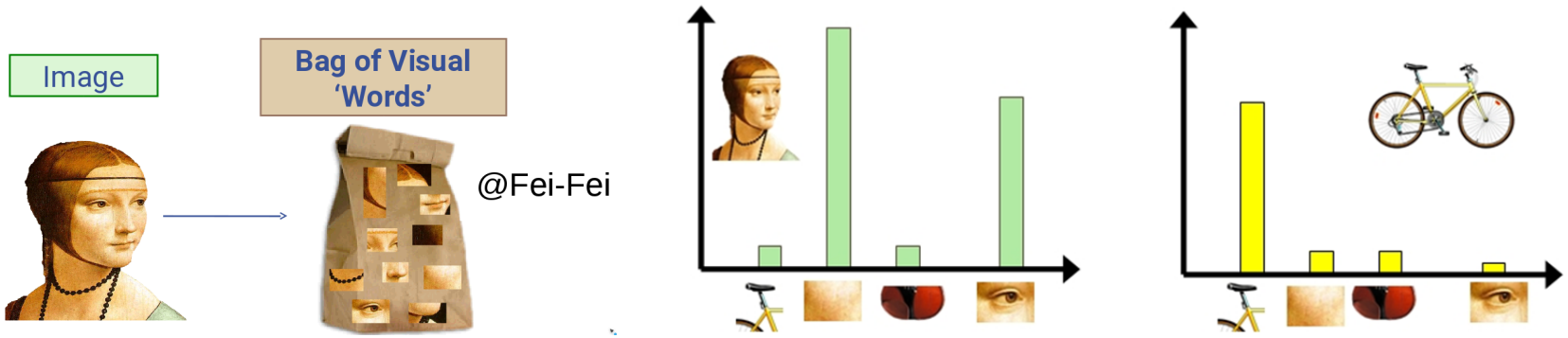


- Bow representation as input for powerful classifiers, *e.g.* SVM

Deep Learning: Trends and methods in the last four decades

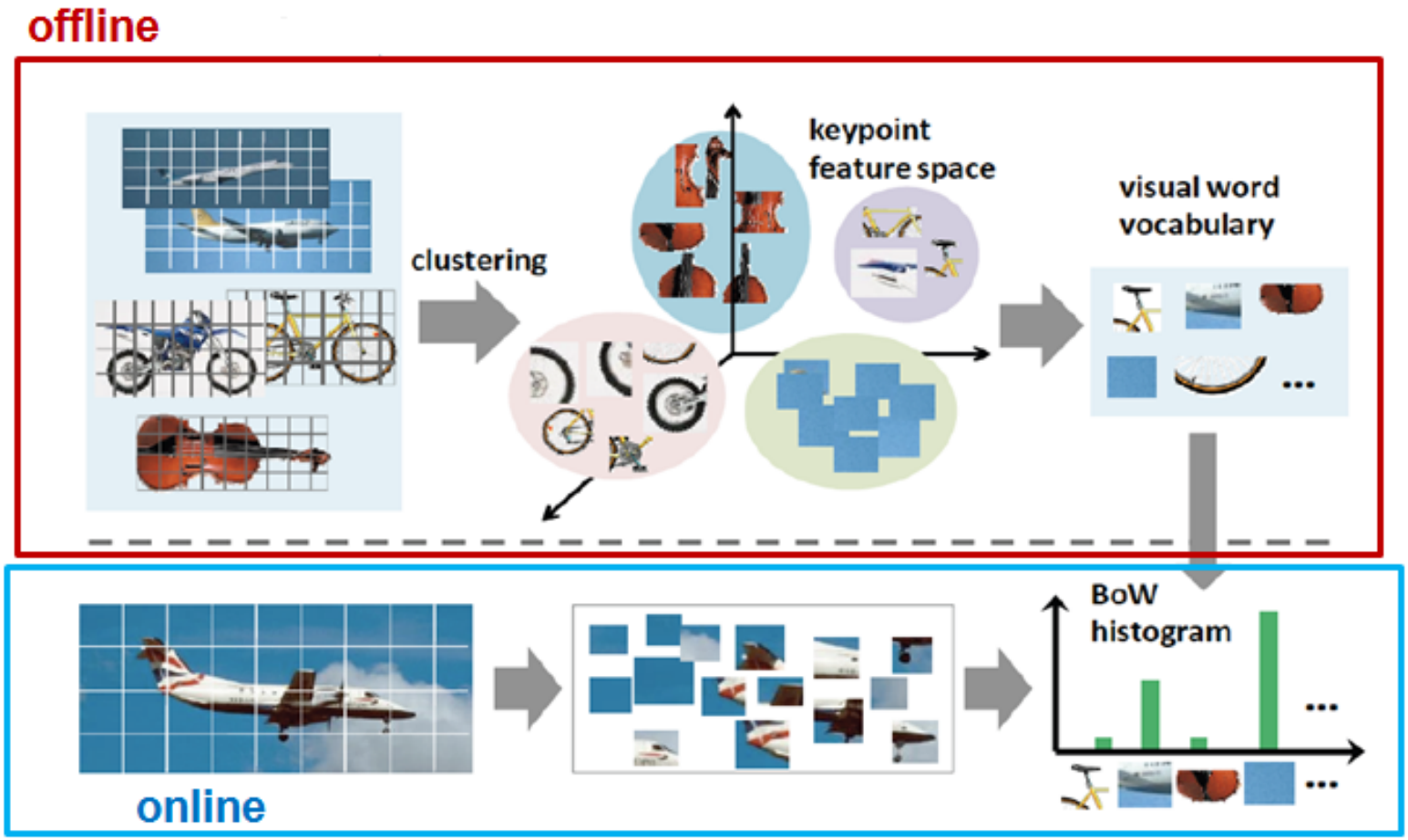
2000's: Bag of Words Model

- Adapting the BoW model for visual recognition ?
  ⇒ Bag of Visual Word (BoV)
- Main challenge: definition of visual words unclear!



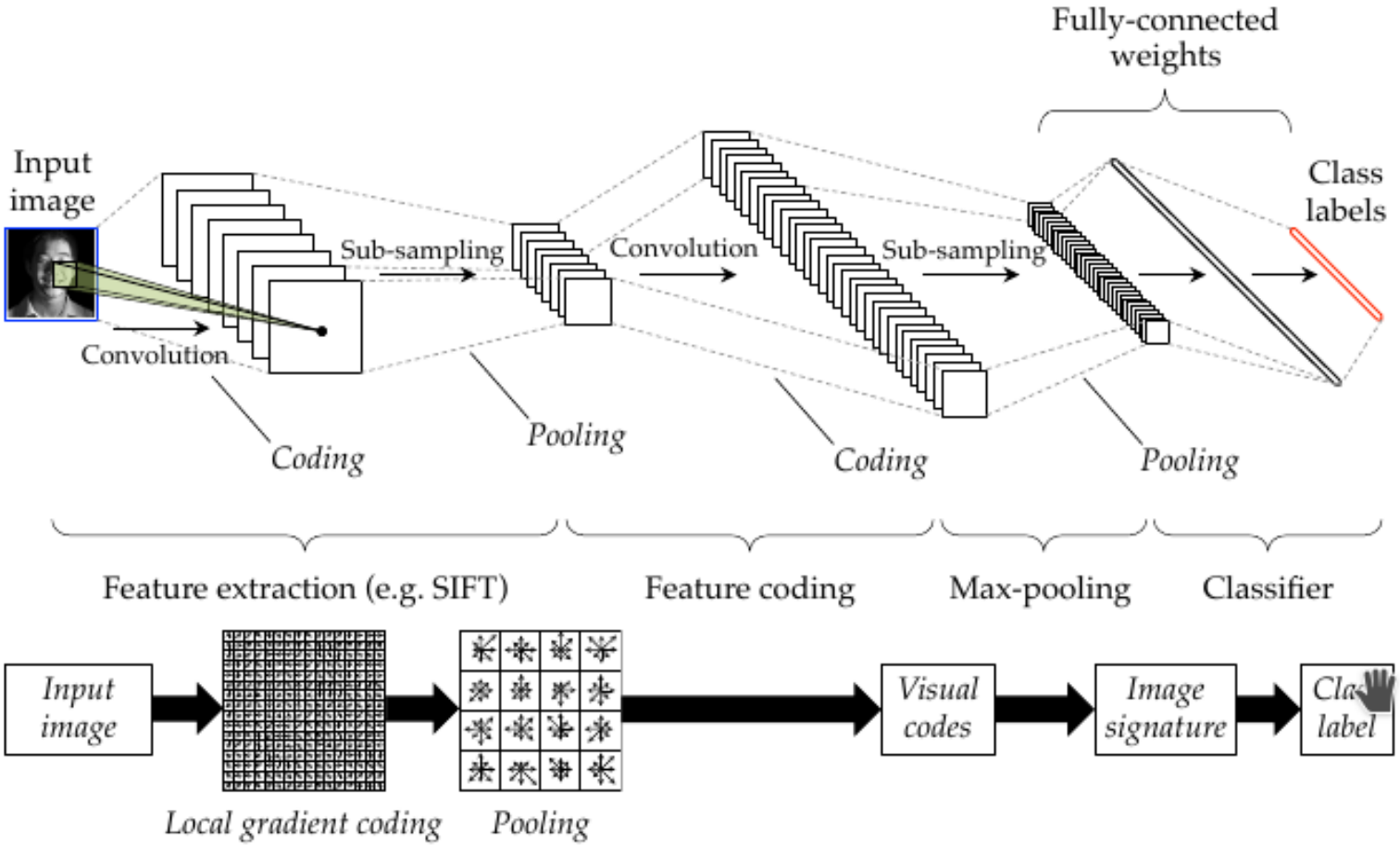
- Solution: compute a dictionary on local image regions (clustering)
  - Local regions represented by handcrafted descriptors, *e.g.* SIFT

2000's: Bag of Visual Words Model



- 2000's: BoW + SVM state-of-the-art
- Many works on kernel on BoW, coding & pooling → 2012

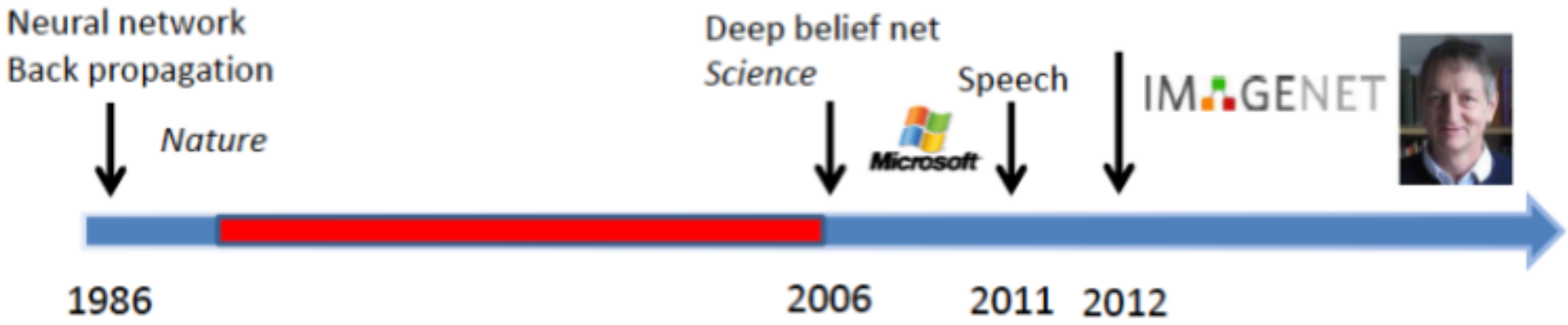
BoW *vs* ConvNet



- BoW architecture: ~ 2 block [Convolution+Pooling] ConvNet !
- ConvNet : learned features, more semantics with larger hierarchies
- **BUT:** not enough training data to learn such models in the 2000's !

Deep Learning: Trends and methods in the last four decades

Deep Learning renewal since 2006

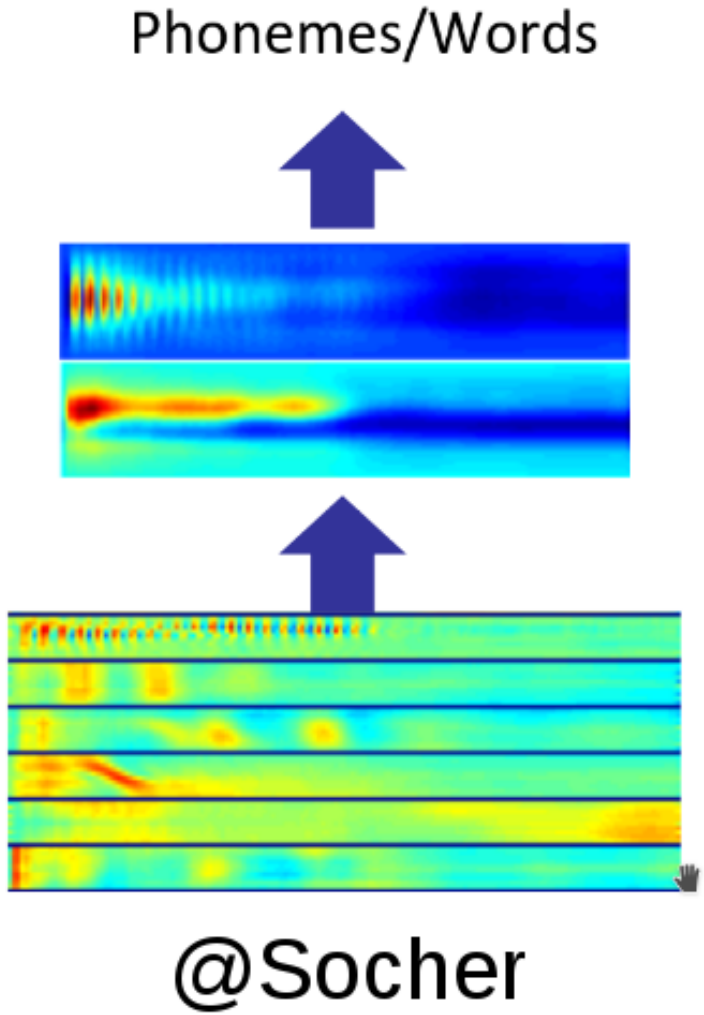


- 2006: new unsupervised learning for Deep Belief Nets (DBN) [HOT06]
- Theoretical results for improving model quality with depth
- Unsupervised training used as init for supervised learning with back-prop

Deep Learning and ConvNet for Speech Recognition

- First DL breakthrough on large datasets: speech recognition
- Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition, Dahl et al. (2010)

Acoustic model	Recog \ WER	RT03S FSH	Hub5 SWB
Traditional features	1-pass –adapt	**27.4**	**23.6**
Deep Learning	1-pass –adapt	**18.5** (–33%)	**16.1** (–32%)



Deep Learning and ConvNet for Image Classification

- ImageNet ILSVRC Challenge (Stanford):
  - $1,200,000$ training images, $1,000$ classes, mono-label
  - Based on WordNet hierarchy (ontology)
  - Evaluation: top-5 error
- Up to 2012, leading approaches: BoW + SVM
- ILSVRC'12: the deep revolution $\Rightarrow$ outstanding success of ConvNets [KSH12]

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	

Outline

1 Deep Learning History

2 2012: the Deep Learning Revolution

2012: the deep revolution

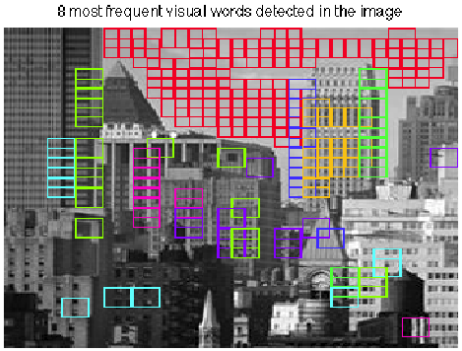
Deep ConvNet success at ILSVRC'12

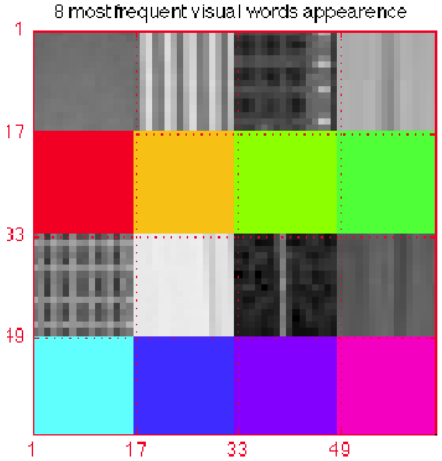
Two main practical reasons:

1. Huge number of labeled images (10^6 images)
   - Possible to train very large models without over-fitting
   - Larger models enables to learn rich (semantic) features hierarchies
2. GPU implementation for training
   - Relatively cheap and fast GPU
   - Training time reduced to 1-2 weeks (up to 50x speed up)

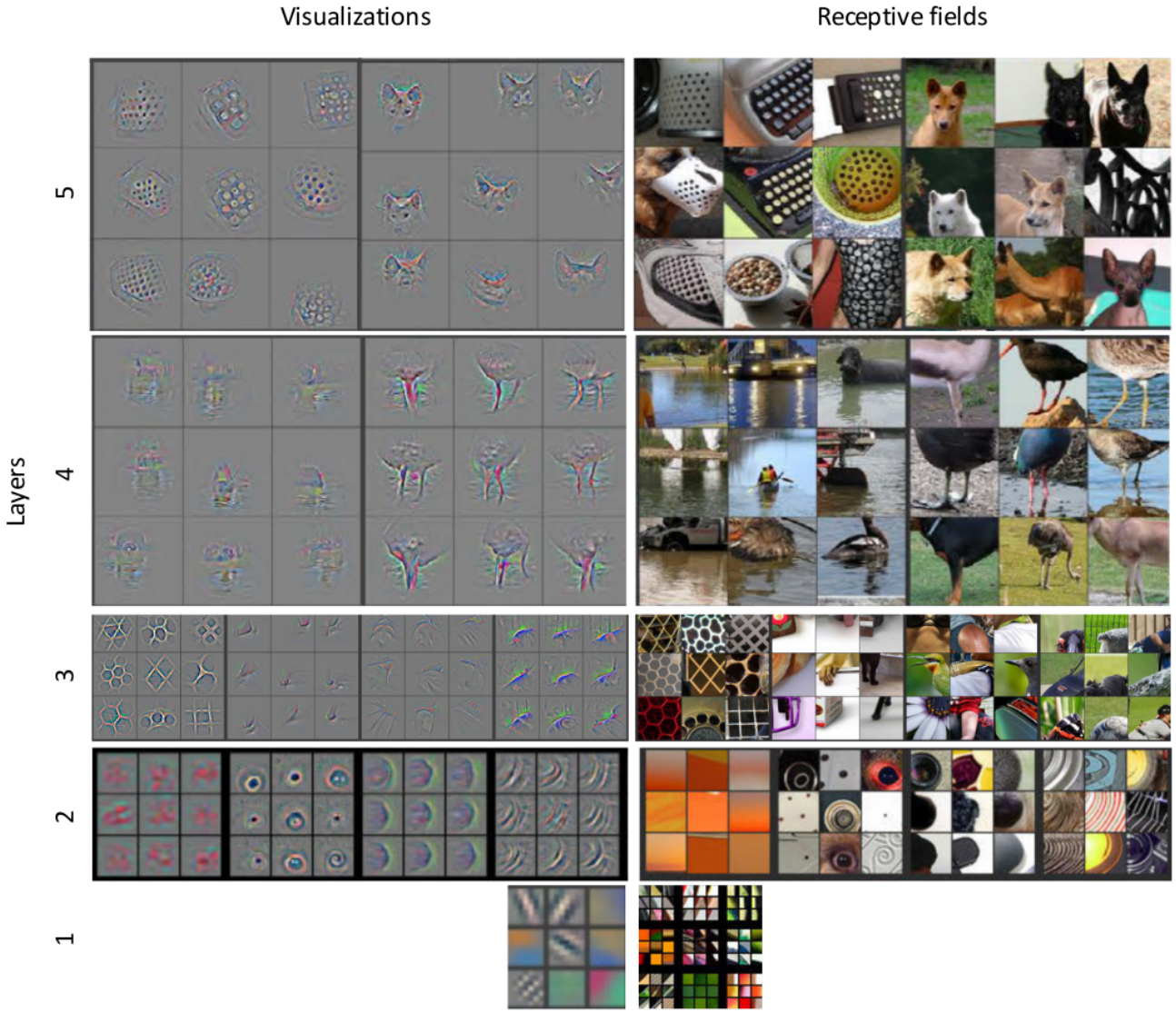
IMAGENET

Deep Learning in 2012: Representation Learning



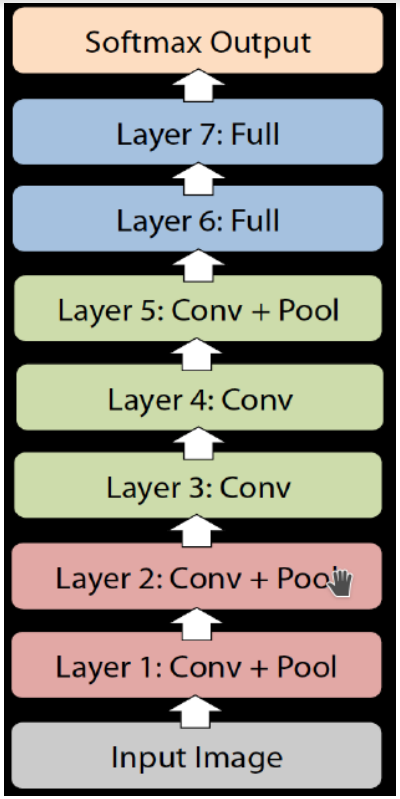


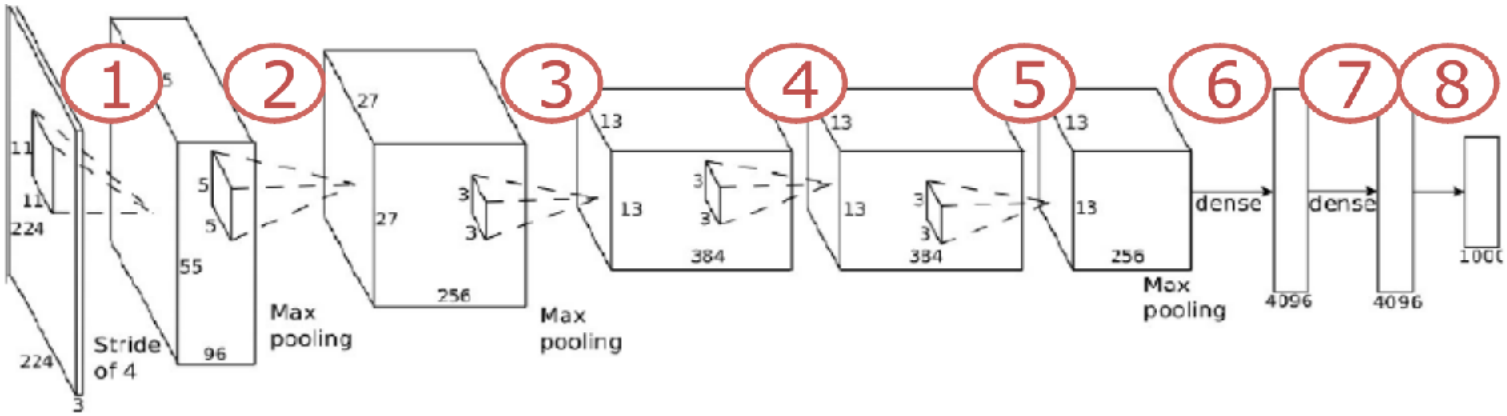
Deep: more semantic features



AlexNet [KSH12] in ILSVRC'12

- 60,000,000 parameters
- 650,000 neurons - 630,000,000 connections
- 5 convolutional layers, 3 Fully Connected (FC)
  - Convolution layer: Convolution + non linearity (ReLU) + pooling
  - Full= FC + non linearity - Final FC: 4096-dim
- Trained on 2 GPUs for a week





AlexNet [KSH12] in ILSVRC'12

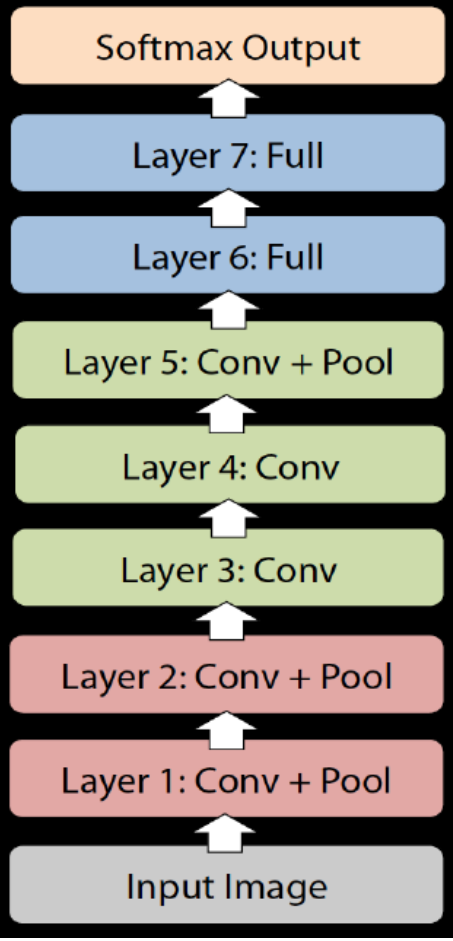
First Convolutionnal Layer

- Input: Images: 227x227x3
- Filter (receptive field) size F: 11, S (stride) = 4
- 96 filters ⇒ output size 55*55*96 = 290,400 neurons
- Each Filter has 11*11*3 = 363 weights + 1 bias = 364 params
  - N.B.: Convolution in whole feature map depth (*cf* LeNet 5 discussion)
- # parms: 96 * 364 = 34, 944

AlexNet [KSH12] in ILSVRC'12

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

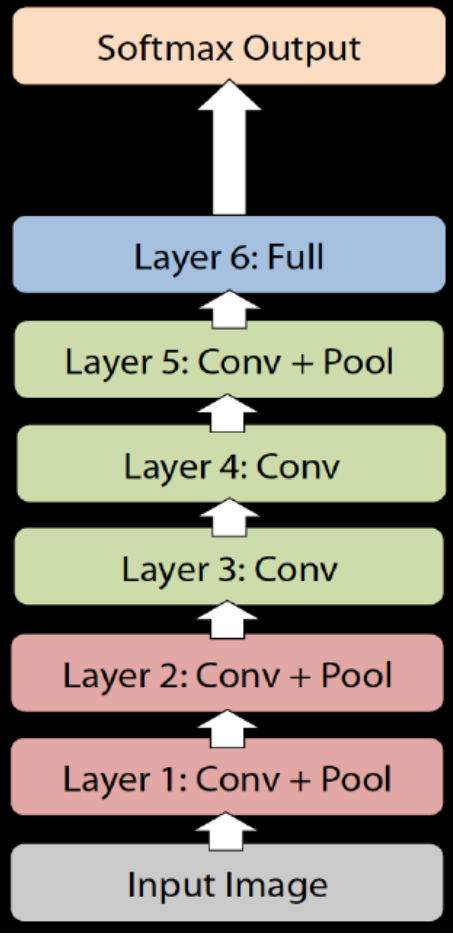


Credit: R. Fergus

AlexNet [KSH12] in ILSVRC'12

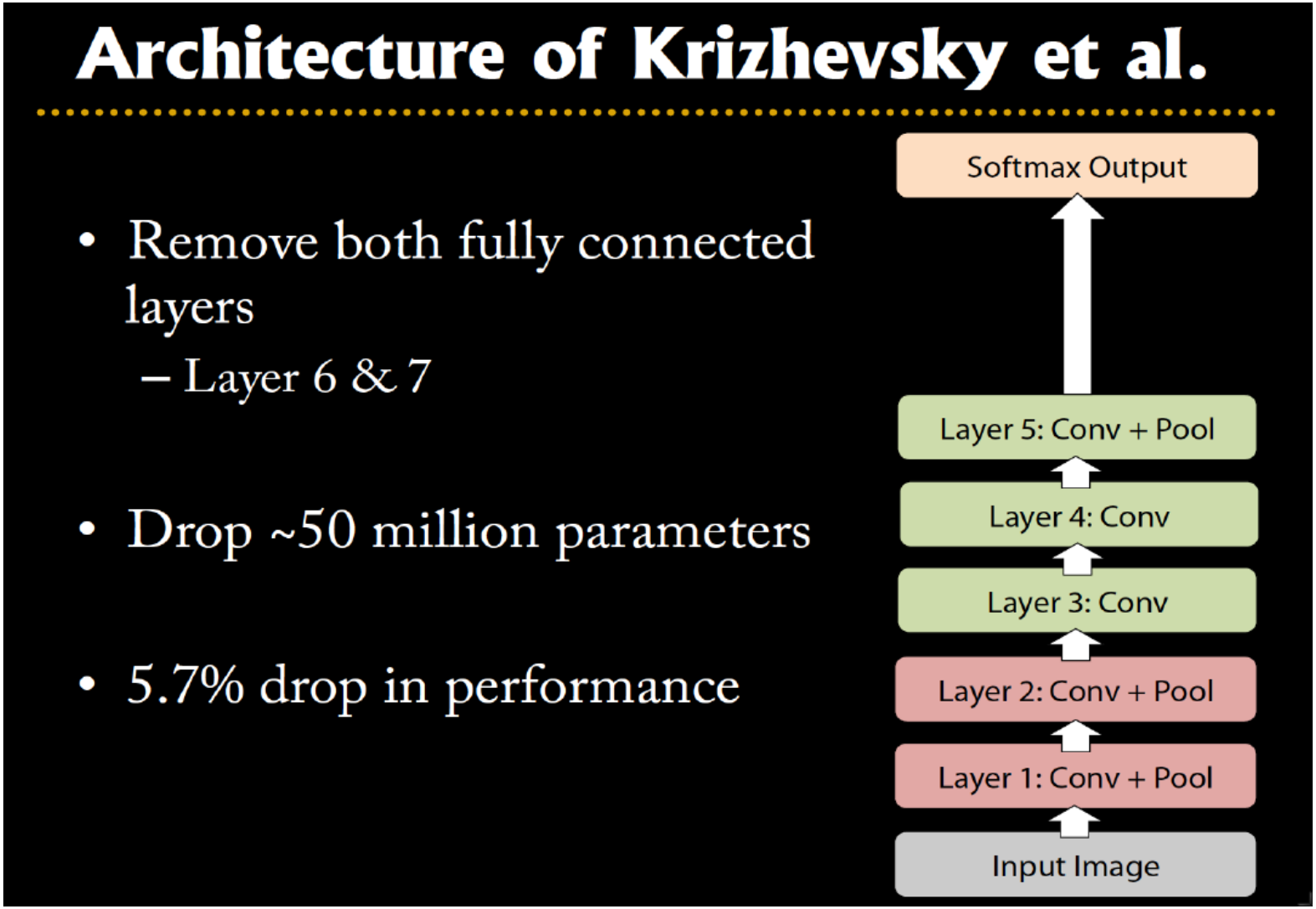
Architecture of Krizhevsky et al.

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



Credit: R. Fergus

AlexNet [KSH12] in ILSVRC'12



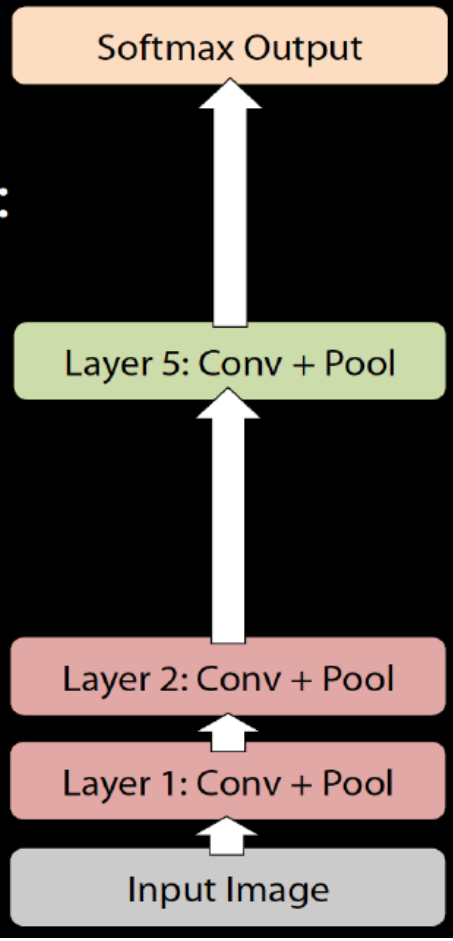
Credit: R. Fergus

AlexNet [KSH12] in ILSVRC'12

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

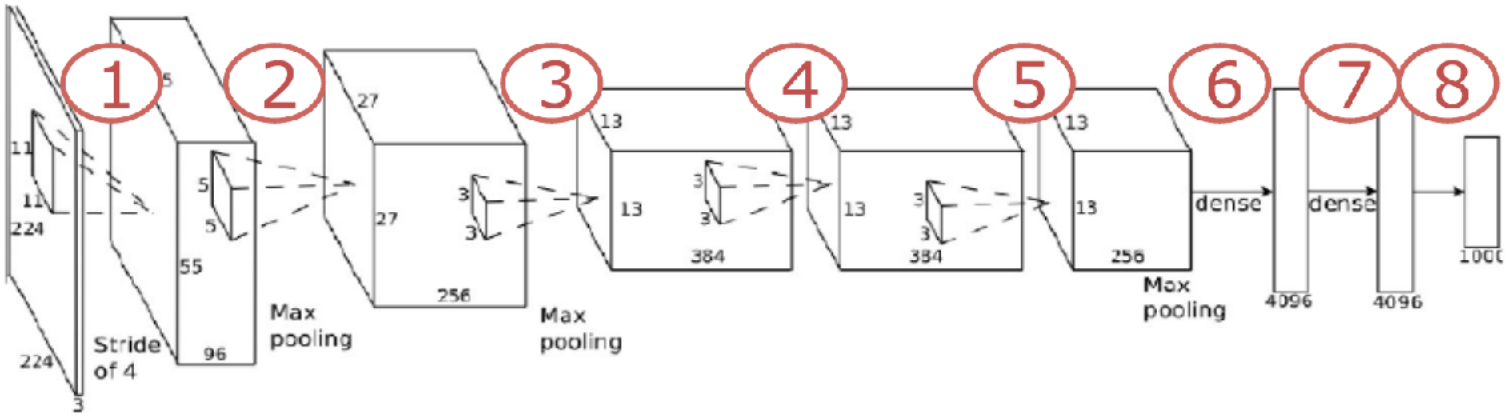
→Depth of network is key



Credit: R. Fergus

AlexNet [KSH12] in ILSVRC'12

- Same global architecture as older nets, *e.g.* LeNet
  - Trained with back-prop and stochastic gradient descent
- But bigger (deeper and wider): 60 10^6 parameters *vs* 60 10^3
  - Needs more data (10^6 *vs* 10^4)
  - GPU implementation for fast training
- Also some architectural and optim improvements (see next course):
  - Non-linearity: ReLU *vs* sigmoid
  - Overlapping pooling (Local Response Normalisation, LRN)
  - Regularization: data augmentation, dropout



References I


Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh, *A fast learning algorithm for deep belief nets*, Neural Comput. 18 (2006), no. 7, 1527–1554.


Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, *Imagenet classification with deep convolutional neural networks*, Advances in neural information processing systems, 2012, pp. 1097–1105.


Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel, *Backpropagation applied to handwritten zip code recognition*, Neural computation 1 (1989), no. 4, 541–551.