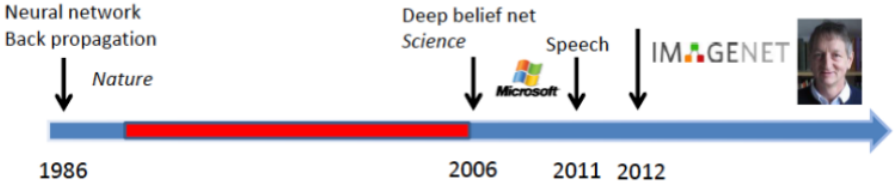
Neural Networks and Deep Learning: Deep Learning Renewal

Nicolas Thome

Conservatoire National des Arts et Métiers (Cnam)
Département Informatique

Deep Learning Renewal since 2006

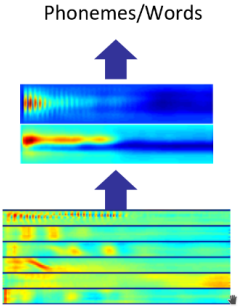


- 2006: new unsupervised learning for Deep Belief Nets (DBN) [Hinton et al., 2006]
- Theoretical results for improving model with depth
- Unsupervised training used as init for back-prop

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., 2012]

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 [Krizhevsky et al., 2012]

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	

2012: the deep revolution

Deep ConvNet success at ILSVRC'12: Two main practical reasons:

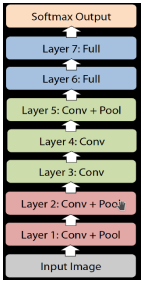
1. Huge number of labeled images (10^6 images)
   - Possible to train very large models without over-fitting
   - Larger models enables to learn rich (semantic) features hierarchies
2. GPU implementation for training
   - Relatively cheap and fast GPU
   - Training time reduced to 1-2 weeks (up to 50x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

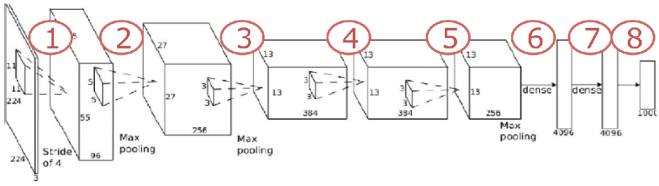
IMAGENET

AlexNet [Krizhevsky et al., 2012] at ILSVRC'12

- 60,000,000 parameters
- 650,000 neurons - 630,000,000 connections
- 5 convolutional layers, 3 Fully Connected (FC)

  $\overbrace{\text{Convolution Layer}}^{\text{convolution + non linearity}}$ + pooling
  - Full= FC + non linearity - Final FC: 4096-dim
- Trained on 2 GPUs for a week

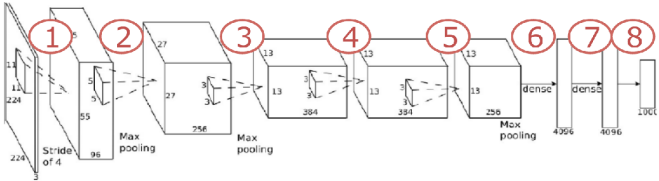




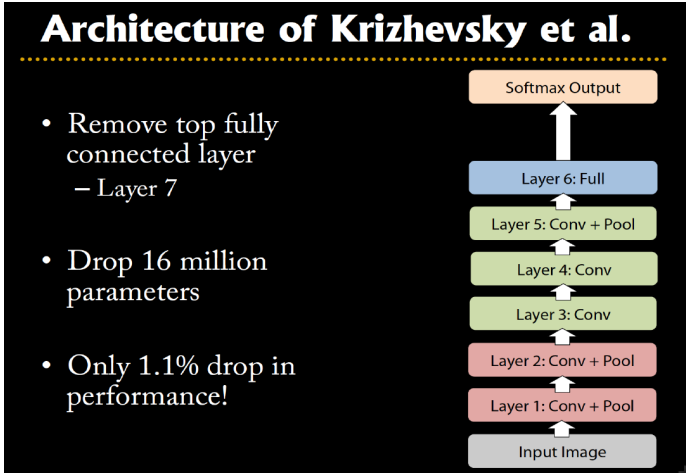
AlexNet [Krizhevsky et al., 2012] at ILSVRC'12

First Convolutionnal Layer

- Input Tensor: $227 \times 227 \times 3$ (color image)
- 96 Filter of size 11x11x3, stride 4
- Output Tensor: 55x55x96 = 290,400 neurons
- Each filter: 11*11*3 +1 = 364 params
  - N.B.: whole feature map convolution (*cf* LeNet5)
- # parms: 96 * 364 = 34, 944

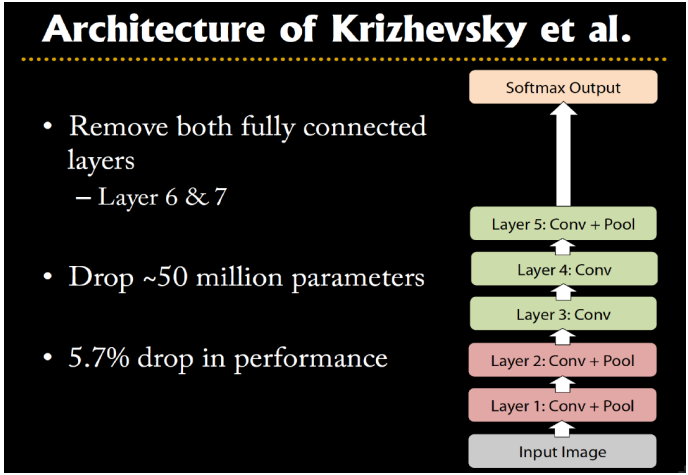


AlexNet [Krizhevsky et al., 2012] at ILSVRC'12



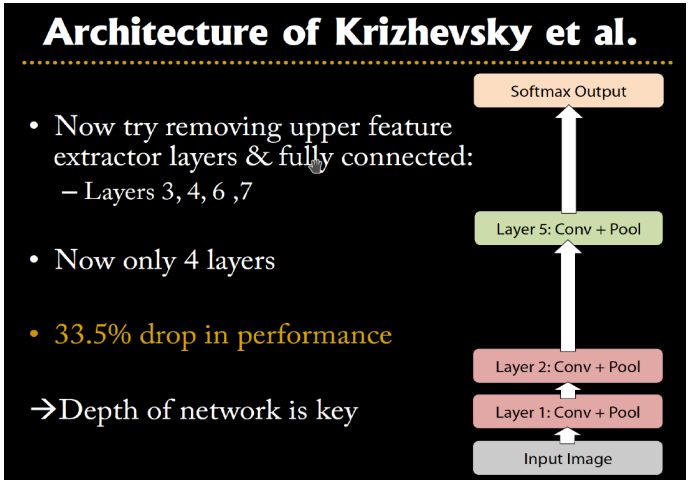
Credit: R. Fergus

AlexNet [Krizhevsky et al., 2012] at ILSVRC'12



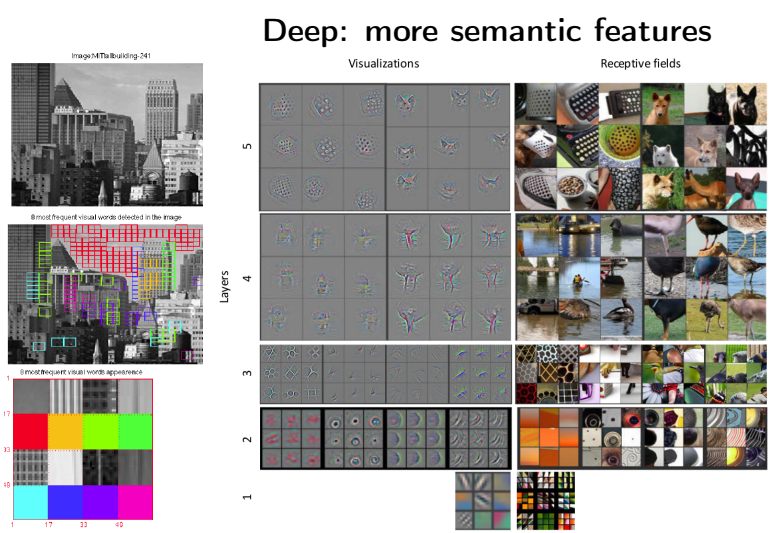
Credit: R. Fergus

AlexNet [Krizhevsky et al., 2012] at ILSVRC'12



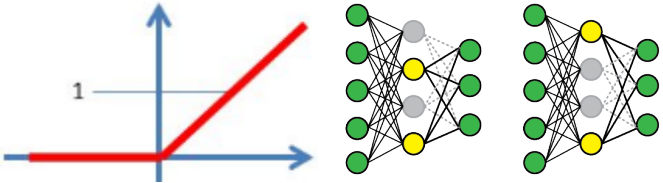
Credit: R. Fergus

Deep Learning in 2012: Representation Learning



AlexNet [Krizhevsky et al., 2012] at ILSVRC'12

- **ILSVRC'12: start of a new era for deep learning**
- AlexNet: macro architecture [Conv-Pool] + FC ~ 80's 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 architectural and optimization improvements**
  $\Rightarrow$ following!



References I

Dahl, G. E., Yu, D., Deng, L., and Acero, A. (2012).
Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition.
Trans. Audio, Speech and Lang. Proc., 20(1):30–42.

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006).
A fast learning algorithm for deep belief nets.
Neural Comput., 18(7):1527–1554.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012).
Imagenet classification with deep convolutional neural networks.
In *Advances in neural information processing systems*, pages 1097–1105.