JHU vision lab

Mathematics of Deep Learning

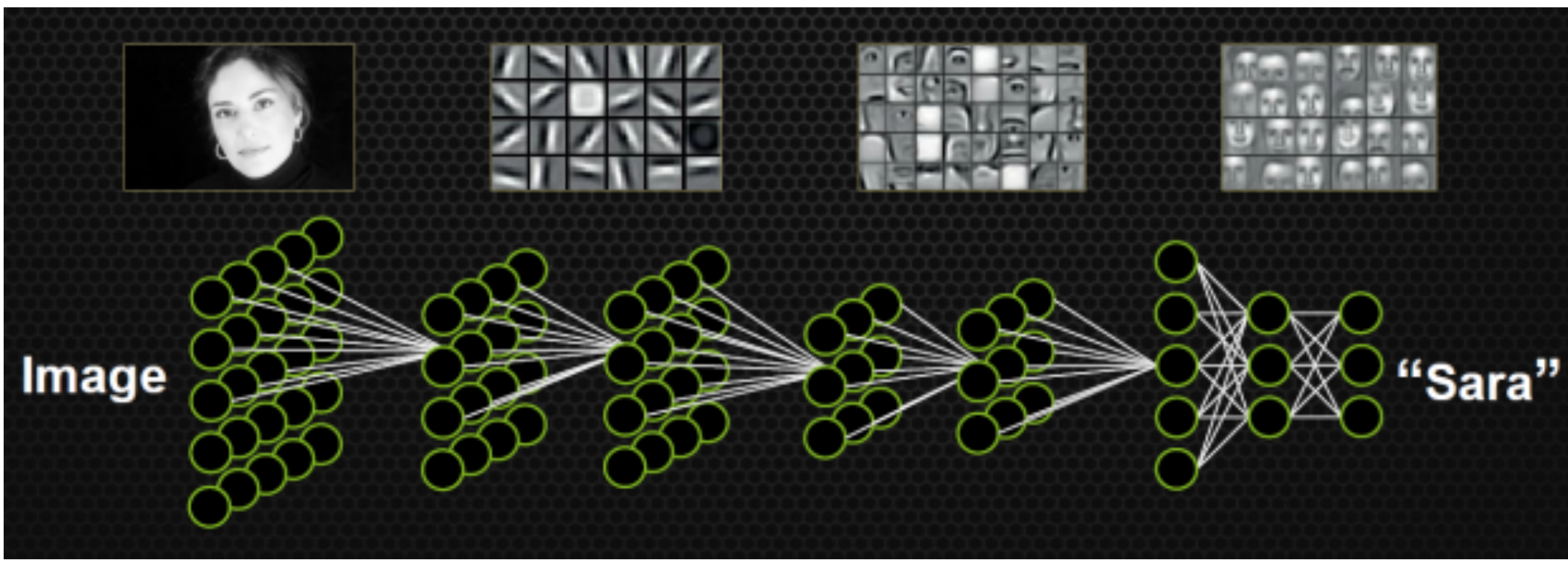
CVPR Tutorial, Honolulu, USA, July 21st 2017

Raja Giryes (Tel Aviv University), **René Vidal** (Hopkins)

Center for IMAGING SCIENCE

Learning Deep Image Feature Hierarchies

- Deep learning gives ~ 10% improvement on ImageNet
  - 1.2M images
  - 1000 categories
  - 60 million parameters



[1] Krizhevsky, Sutskever and Hinton. ImageNet classification with deep convolutional neural networks, NIPS’12.
[2] Sermanet, Eigen, Zhang, Mathieu, Fergus, LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. ICLR’14.
[3] Donahue, Jia, Vinyals, Hoffman, Zhang, Tzeng, Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. ICML’14.

Impact of Deep Learning in Computer Vision

- 2012-2014 classification results in ImageNet

CNN
non-CNN

2012 Teams	%error
Supervision (Toronto)	15.3
ISI (Tokyo)	26.1
VGG (Oxford)	26.9
XRCE/INRIA	27.0
UvA (Amsterdam)	29.6
INRIA/LEAR	33.4

2013 Teams	%error
Clarifai (NYU spinoff)	11.7
NUS (singapore)	12.9
Zeiler-Fergus (NYU)	13.5
A. Howard	13.5
OverFeat (NYU)	14.1
UvA (Amsterdam)	14.2
Adobe	15.2
VGG (Oxford)	15.2
VGG (Oxford)	23.0

2014 Teams	%error
GoogLeNet	6.6
VGG (Oxford)	7.3
MSRA	8.0
A. Howard	8.1
DeeperVision	9.5
NUS-BST	9.7
TTIC-ECP	10.2
XYZ	11.2
UvA	12.1

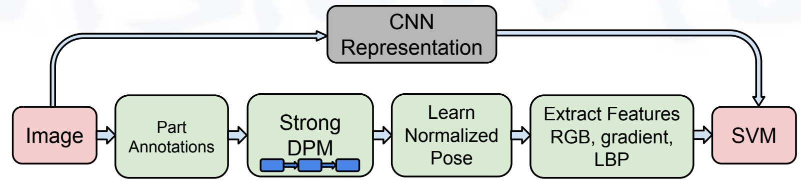
- 2015 results: MSR under 3.5% error using 150 layers!

Slide from Yann LeCun's CVPR'15 plenary and ICCV'15 tutorial intro by Joan Bruna

Transfer from ImageNet to Other Datasets

- CNNs + SMVs [1]

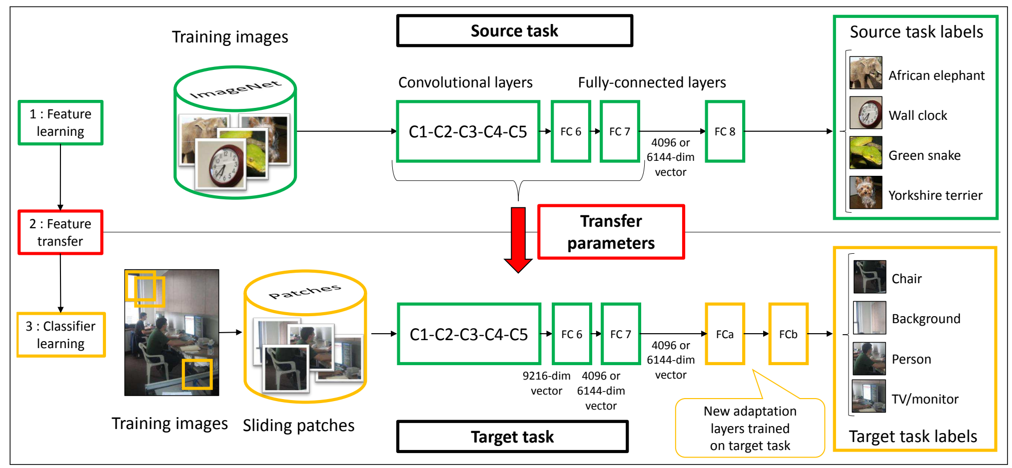
Pascal VOC 2007	mAP
GHM[8]	64.7
AGS[11]	71.1
NUS[39]	70.5
CNN-SVM	73.9
CNNaug-SVM	**77.2**



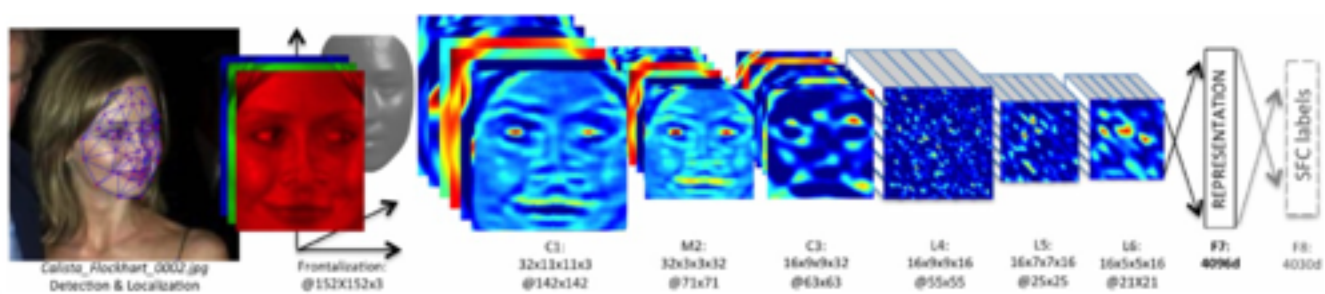
- Retrain top-layer [2]

Pascal VOC 2007	mAP
INRIA [32]	59.4
NUS-PSL [44]	70.5
PRE-1000C	**77.7**

Pascal VOC 2012	mAP
NUS-PSL [49]	82.2
NO PRETRAIN	70.9
PRE-1000C	78.7
PRE-1000R	76.3
PRE-1512	**82.8**



- Deep Face [3]

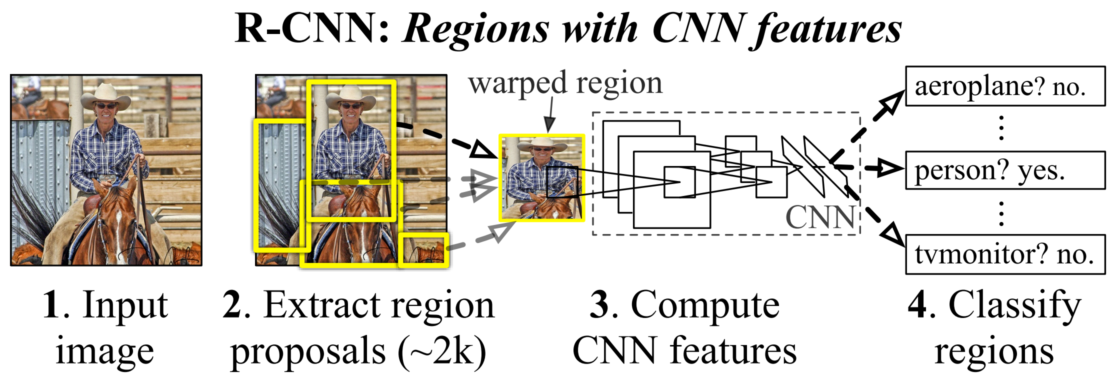


[1] Razavian, Azizpour, Sullivan, Carlsson, CNN Features off-the-shelf: an Astounding Baseline for Recognition. CVPRW'14.
[2] Oquab, Bottou, Laptev, Sivic. Learning and transferring mid-level image representations using convolutional neural networks CVPR'14
[3] Taigman, Yang, Ranzato, Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. CVPR'14

Transfer from Classification to Other Tasks

- CNNs + SVMs for object detection [1,2]

VOC 2010 test	mAP
DPM v5 [20]†	33.4
UVA [39]	35.1
Regionlets [41]	39.7
SegDPM [18]†	40.4
R-CNN	50.2
R-CNN BB	**53.7**

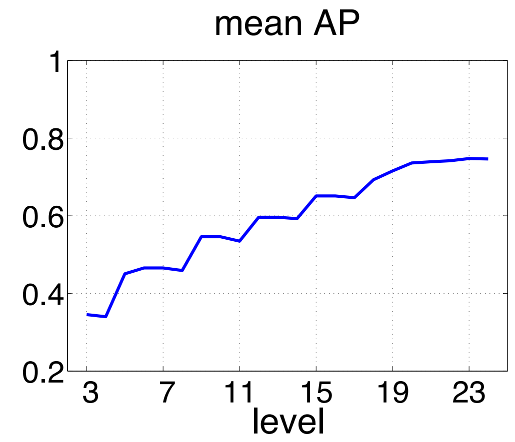


- CNNs for pose estimation [3] and semantic segmentation [4]

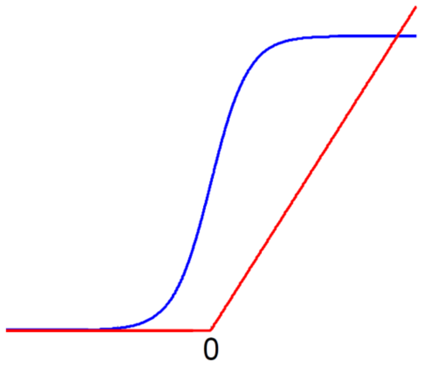

[1] Girshick, Donahue, Darrell and Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, CVPR'14
[2] Sermanet, Eigen, Zhang, Mathieu, Fergus, LeCun. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. ICLR
[3] Tompson, Goroshin, Jain, LeCun, Bregler. Efficient Object Localization Using Convolutional Networks. CVPR'15
[4] Pinheiro, Collobert, Dollar. Learning to Segment Object Candidates. NIPS'15

Why These Improvements in Performance?

- Features are learned rather than hand-crafted
- More layers capture more invariances [1]



- More data to train deeper networks
- More computing (GPUs)
- Better regularization: Dropout
- New nonlinearities
  - Max pooling, Rectified linear units (ReLU)



- Theoretical understanding of deep networks remains shallow

[1] Razavian, Azizpour, Sullivan, Carlsson, CNN Features off-the-shelf: an Astounding Baseline for Recognition. CVPRW'14.

Theoretical Results on Deep Learning

- **Approximation, depth, width, and invariance theory**
  - Perceptrons and multilayer feedforward networks are universal approximators: Cybenko ’89, Hornik ’89, Hornik ’91, Barron ’93
  - Scattering networks are deformation stable for Lipschitz non-linearities: Bruna-Mallat ’13, Wiatowski ’15, Mallat ’16

- **Generalization and regularization theory**
  - # training examples grows exponentially with network size: Barlett ‘03
  - Distance and margin-preserving embeddings: Giryes ’15, Sokolik ‘16
  - Geometry, generalization bounds and depth efficiency: Montufar ’15, Neyshabur ’15, Shashua ’14 ’15 ‘16

[1] Cybenko. Approximations by superpositions of sigmoidal functions, Mathematics of Control, Signals, and Systems, 2 (4), 303-314, 1989.
[2] Hornik, Stinchcombe and White. Multilayer feedforward networks are universal approximators, Neural Networks, 2(3), 359-366, 1989.
[3] Hornik. Approximation Capabilities of Multilayer Feedforward Networks, Neural Networks, 4(2), 251–257, 1991.
[4] Barron. Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information Theory, 39(3):930–945, 1993.
[5] Bruna and Mallat. Invariant scattering convolution networks. Trans. PAMI, 35(8):1872–1886, 2013.
[6] Wiatowski, Bölcskei. A mathematical theory of deep convolutional neural networks for feature extraction. arXiv 2015.
[7] Mallat. Understanding deep convolutional networks. Phil. Trans. R. Soc. A, 374(2065), 2016.

[8] Bartlett and Maass. Vapnik-Chervonenkis dimension of neural nets. The handbook of brain theory and neural networks, pages 1188– 1192, 2003.
[9] Giryes, Sapiro, A Bronstein. Deep Neural Networks with Random Gaussian Weights: A Universal Classification Strategy? arXiv:1504.08291.
[10] Sokolic. Margin Preservation of Deep Neural Networks, 2015
[11] Montufar. Geometric and Combinatorial Perspectives on Deep Neural Networks, 2015.
[12] Neyshabur. The Geometry of Optimization and Generalization in Neural Networks: A Path-based Approach, 2015.

Theoretical Results on Deep Learning

- **Earlier work on optimization theory**
  - No spurious local optima for linear networks (Baldi & Hornik ’89)
  - Backpropagation fails to converge for nonlinear networks (Brady ’89)
  - Back propagation converges for linearly separable data (Gori & Tesi ’91 ’92), but it get stuck in other cases (Frasconi ’97)

- **Recent work on optimization theory**
  - Convex neural networks in infinite number of variables: Bengio ‘05
  - Networks with many hidden units can learn polynomials: Andoni‘14
  - The loss surface of multilayer networks: Choromanska ’15
  - Attacking the saddle point problem: Dauphin ‘14
  - Effect of gradient noise on the energy landscape: Chaudhari ‘15
  - Guaranteed training of NNs using tensor methods: Janzamin ’15
  - Guarantees of global optimality in neural network training: Haeffele ‘15

Motivations and Goals of this Tutorial

- **Motivation:** Deep networks have led to dramatic improvements in performance for many tasks, but the mathematical reasons for this success remain unclear.

- **Goal:** Review very recent work that aims at understanding the mathematical reasons for the success of deep networks.

- **What we will do:** Study theoretical questions such as
  - Can we ensure that the learned representations are globally optimal?
  - What properties of images are being captured/exploited by DNNs?
  - How should DNNs be regularized to ensure generalization properties?

- **What we will not do:** Show X% improvement in performance for a particular application.

Tutorial Schedule

- 09:00-09:15: **René Vidal** - Introduction
- 09:15-10:00: **René Vidal** - Global Optimality and Regularization in Deep Learning
- 10:00-10:30: **Coffee Break**
- 10:30-11:15: **Raja Giryes** - Data Structure Based Theory of Deep Learning
- 11:15-12:00: **Raja Giryes**- Generalization Bounds for Deep Learning
- 12:00-12:15: **Discussion**