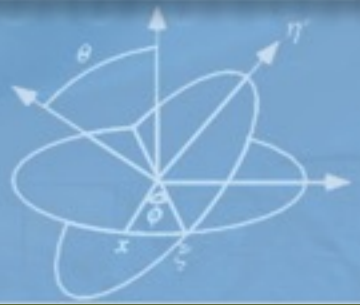
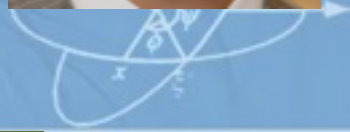
JHU vision lab

Mathematics of Deep Learning

ICCV Tutorial, Venice, Italy, October 22nd, 2017

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



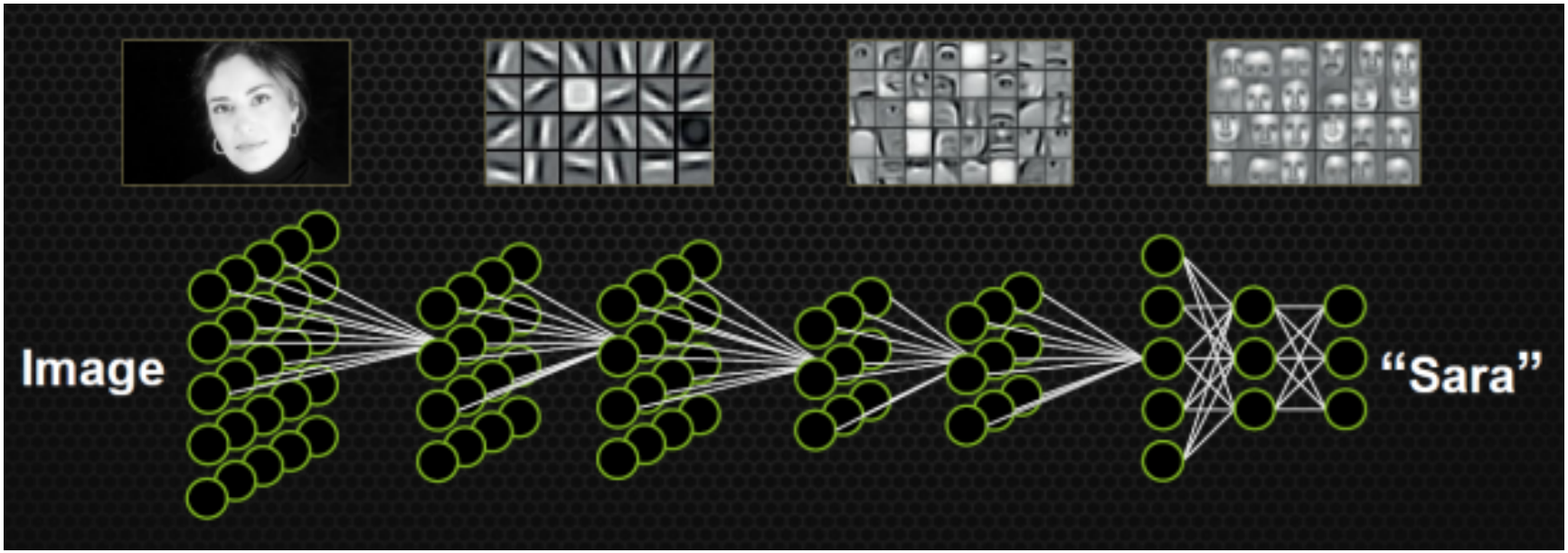


THE DEPARTMENT OF BIOMEDICAL ENGINEERING
The Whitaker Institute at Johns Hopkins

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

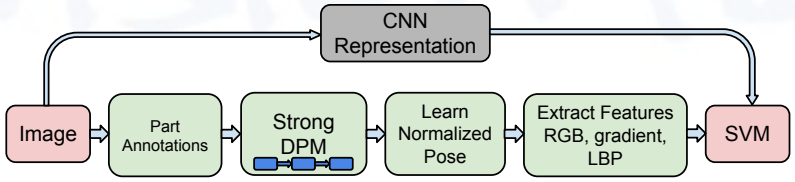
- 2015 results: MSR under 3.5% error using 150 layers!

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]

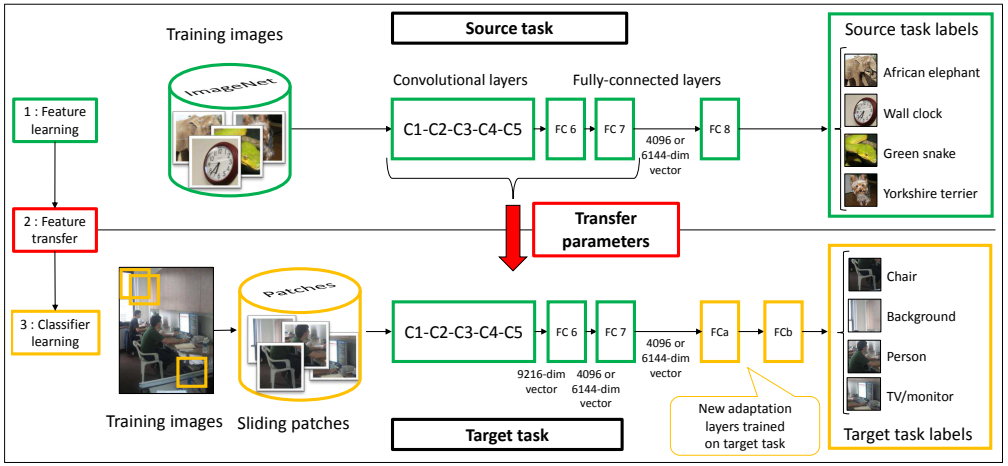
Pascal VOC 2007	mAP
GHM[8]	64.7
AGS[11]	71.1
NUS[39]	70.5
CNN-SVM	73.9
CNNaug-SVM	**77.2**



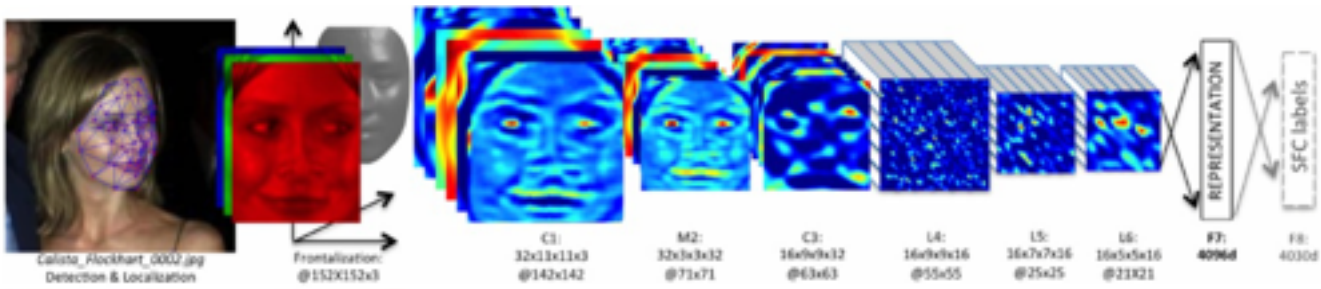
- Retrain top-layer [2]

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]

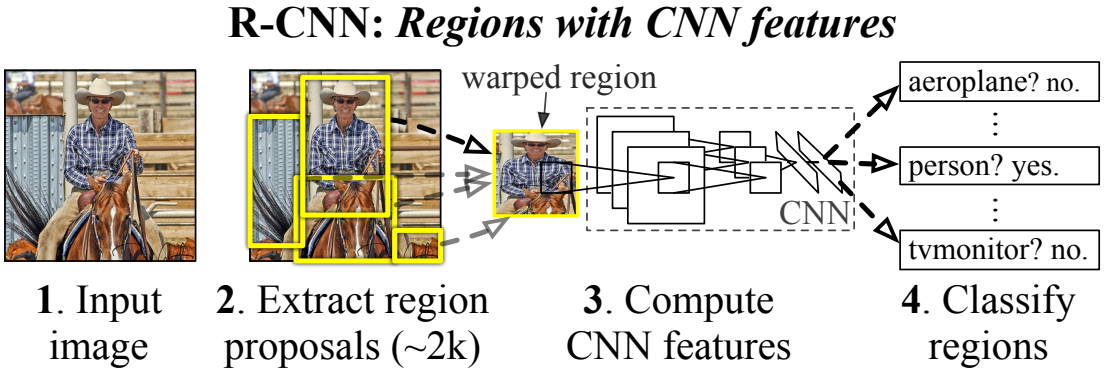


[1] Razavian, Azizpour, Sullivan, Carlsson, CNN Features off-the-shelf: an Astounding Baseline for Recognition. CVPRW'14.
[2] Oquab, Bottou, Laptev, Sivic. Learning and transferring mid-level image representations using convolutional neural networks CVPR'14
[3] Taigman, Yang, Ranzato, Wolf. DeepFace: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

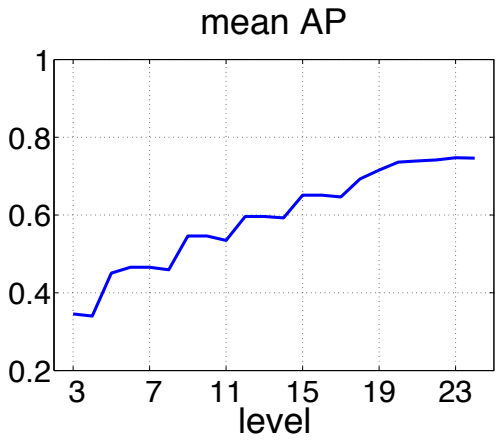


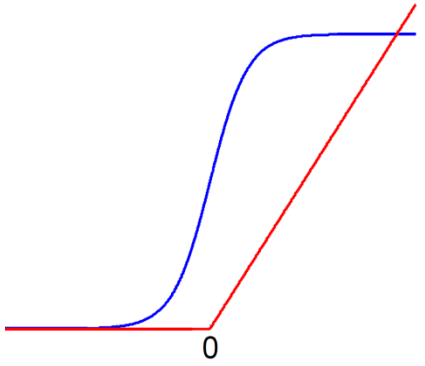
- CNNs for pose estimation [3] and semantic segmentation [4]

[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

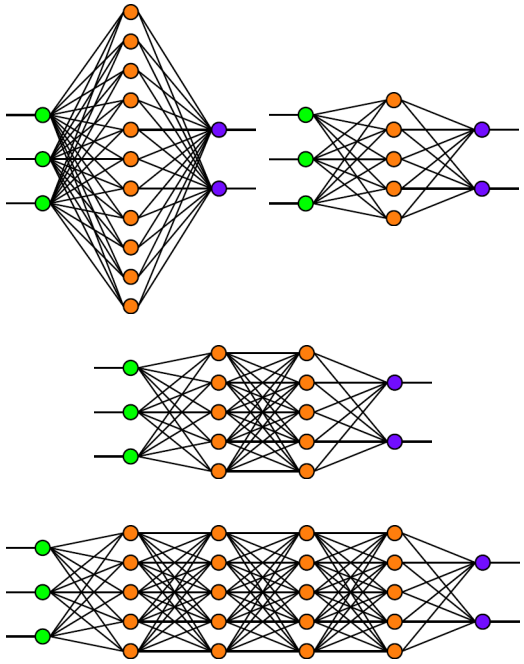




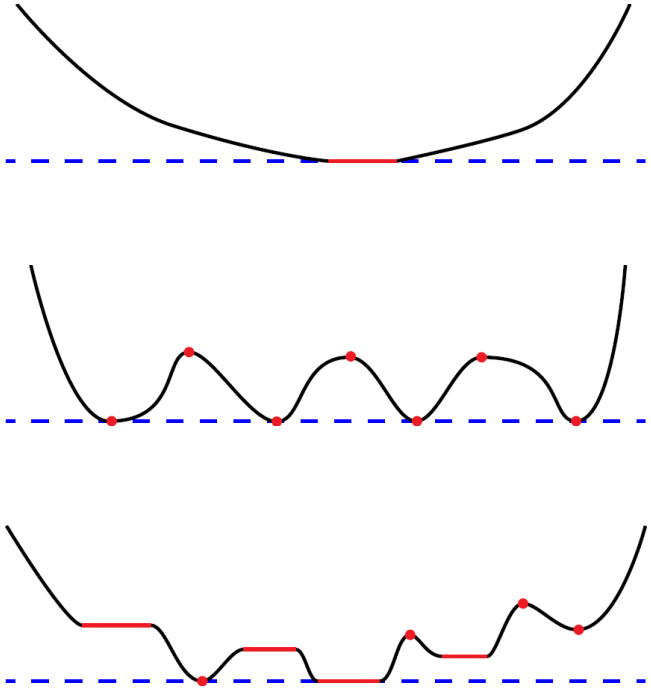
[1] Razavian, Azizpour, Sullivan, Carlsson, CNN Features off-the-shelf: an Astounding Baseline for Recognition. CVPRW'14.

Key Theoretical Questions

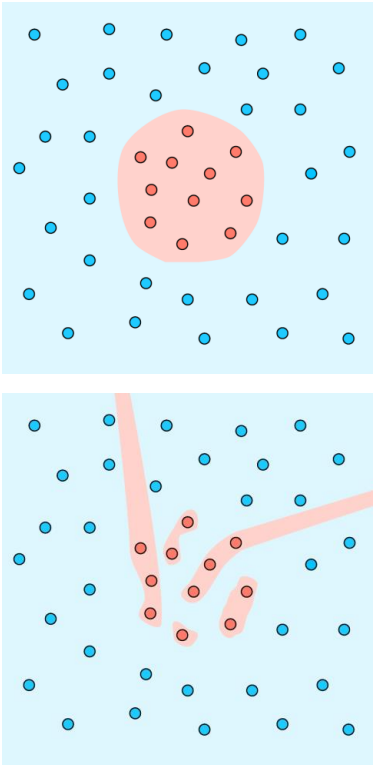
Architecture Design



Optimization



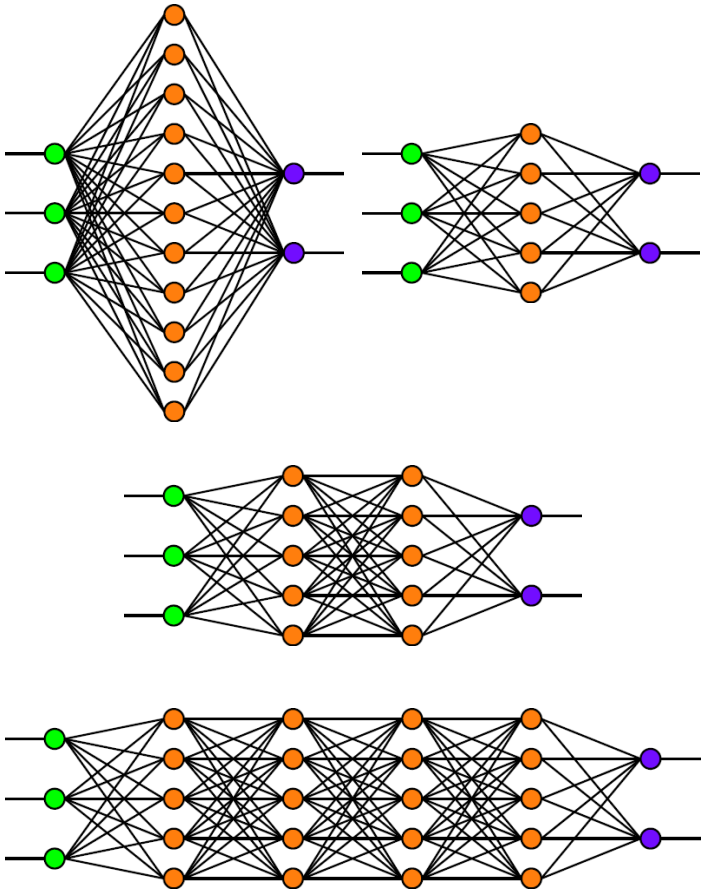
Generalization



Slide courtesy of Ben Haeffele

Key Theoretical Questions: Architecture

- **Are there principled ways to design networks?**
  - How many layers?
  - Size of layers?
  - Choice of layer types?
  - What classes of functions can be approximated by a feedforward neural network?
  - How does the architecture impact expressiveness? [1]



Slide courtesy of Ben Haeffele

[1] Cohen, et al., “On the expressive power of deep learning: A tensor analysis.” COLT. (2016)

Key Theoretical Questions: Architecture

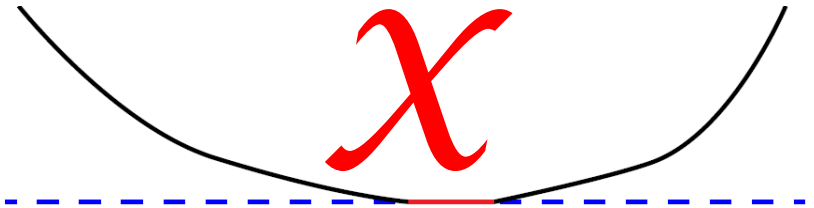
- **Approximation, depth, width and invariance: earlier work**
  - Perceptrons and multilayer feedforward networks are universal approximators: Cybenko ’89, Hornik ’89, Hornik ’91, Barron ’93

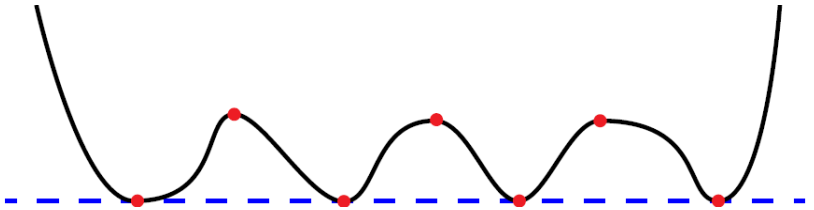
- **Approximation, depth, width and invariance: recent work**
  - Exponential gaps between deep and shallow feedforward networks: Montufar’14
  - Deep narrow Boltzmann machines are universal approximators: Montufar’15
  - Design of CNNs via hierarchical tensor decompositions: Cohen ’17
  - Scattering networks are deformation stable for Lipschitz non-linearities: Bruna-Mallat ’13, Wiatowski ’15, Mallat ’16

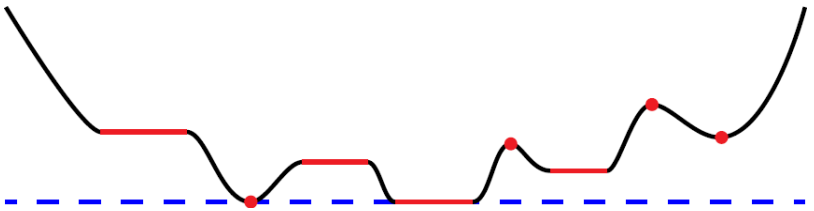
[1] Cybenko. Approximations by superpositions of sigmoidal functions, Mathematics of Control, Signals, and Systems, 2 (4), 303-314, 1989.
[2] Hornik, Stinchcombe and White. Multilayer feedforward networks are universal approximators, Neural Networks, 2(3), 359-366, 1989.
[3] Hornik. Approximation Capabilities 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] Cohen et al. Analysis and Design of Convolutional Networks via Hierarchical Tensor Decompositions arXiv preprint arXiv:1705.02302
[6] Montúfar, Pascanu, Cho, Bengio, On the number of linear regions of deep neural networks, NIPS 27, pp. 2924-2932, 2014
[7] Montúfar et al, Deep narrow Boltzmann machines are universal approximators, ICLR 2015, arXiv:1411.3784v3
[8] Bruna and Mallat. Invariant scattering convolution networks. Trans. PAMI, 35(8):1872–1886, 2013.
[9] Wiatowski, Bölcskei. A mathematical theory of deep convolutional neural networks for feature extraction. arXiv 2015.
[10] Mallat. Understanding deep convolutional networks. Phil. Trans. R. Soc. A, 374(2065), 2016.

Key Theoretical Questions: Optimization

- **How to train neural networks?**
  - Problem is non-convex
  - What does the error surface look like?
  - How to guarantee optimality?
  - When does local descent succeed?







Slide courtesy of Ben Haeffele

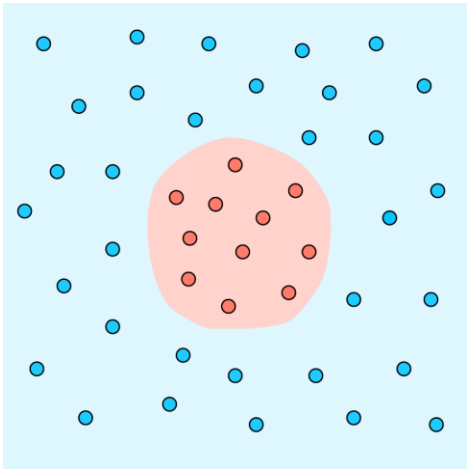
Key Theoretical Questions: Optimization

- **Optimization theory: earlier work**
  - No spurious local minima 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)
- **Optimization theory: recent work**
  - 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
  - Entropy-SGD is biased toward wide valleys: Chaudhari ‘17
  - Deep relaxation: PDEs for optimizing deep nets: Chaudhari ‘17
  - Guaranteed training of NNs using tensor methods: Janzamin ’15
  - No spurious local minima for wide enough networks: Haeffele ‘15

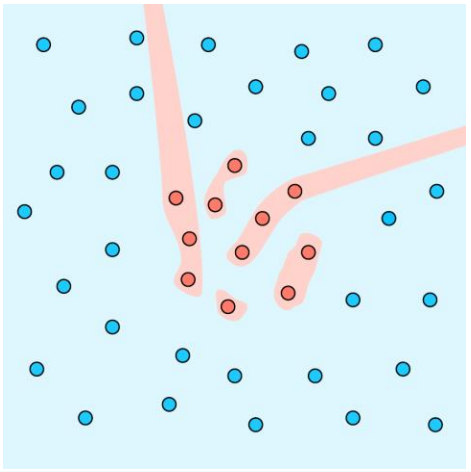
Key Theoretical Questions: Generalization

- **Classification performance guarantees?**
  - How well do deep networks generalize?
  - How should networks be regularized?
  - How to prevent overfitting?

✓ Simple



✗ Complex



Slide courtesy of Ben Haeffele

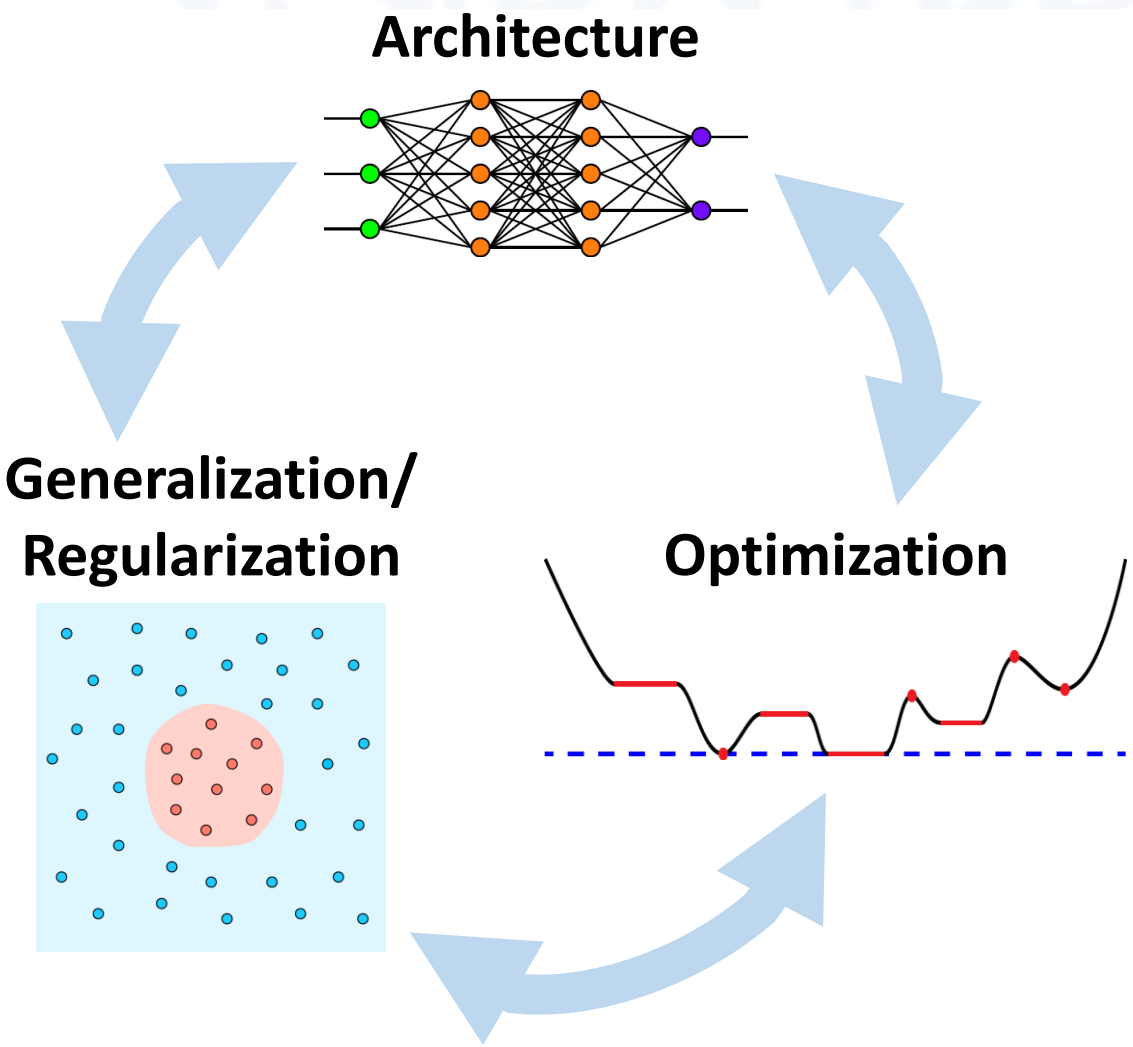
Key Theoretical Questions: Generalization

- **Generalization and regularization theory: earlier work**
  - # training examples grows exponentially with network size [1]
- **New regularization methods**
  - Early stopping [2]
  - Dropout, Dropconnect, and extensions (adaptive, annealed) [3,4]
- **Generalization and regularization theory: recent work**
  - Distance and margin-preserving embeddings [5,6]
  - Path SGD regularization & generalization bounds [7]
  - Product of norms regularization & generalization bounds [8]
  - Implicit regularization & generalization bounds [9]
  - Information theory: information bottleneck, information dropout [10,11]

[1] Bartlett and Maass. Vapnik-Chervonenkis dimension of neural nets. The handbook of brain theory and neural networks, pages 1188– 1192, 2003.
[2] R Caruana, S Lawrence, CL Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. NIPS 2001.
[3] Srivastava. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 2014.
[4] Wan. Regularization of neural networks using dropconnect. In ICML, 2013.
[5] Giryes, Sapiro, A Bronstein. Deep Neural Networks with Random Gaussian Weights: A Universal Classification Strategy? arXiv:1504.08291.
[6] Sokolic. Margin Preservation of Deep Neural Networks, 2015
[7] B Neyshabur. Path-SGD: Path-Normalized Optimization in Deep Neural Networks. NIPS 2015
[8] Sokolić, R. Giryes, G. Sapiro, and M. Rodrigues. Generalization error of invariant classifiers. In AISTATS, 2017.
[9] Behnam Neyshabur. Implicit Regularization in Deep Learning. PhD Thesis 2017
[10] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. arXiv preprint arXiv:1703.00810, 2017.
[11] A. Achille and S. Soatto. Information dropout: Learning optimal representations through noisy computation. *arXiv:1611.01353*, 2016.

Key Theoretical Questions are Interrelated

- Optimization can impact generalization [1]
- Architecture has strong effect on generalization [2]
- Some architectures could be easier to optimize than others



Courtesy of Ben Haeffele

[1] Neyshabur, et al., “In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning.” ICLR workshop. (2015).
[2] Zhang, et al., “Understanding deep learning requires rethinking generalization.” ICLR. (2017).

ICCV 2017 Tutorial Schedule

- 08:30-08:45: **René Vidal** Introduction
- 08:45-09:30: **René Vidal** Global Optimality in Deep Learning
- 09:30-10:15: **Raja Giryes** Structure Based Theory for Deep Learning
- 10:15-11:00: **Coffee Break**
- 11:00-11:45: **Raja Giryes** Generalization Bounds for Deep Learning
- 11:45-12:30: **Vardan Papyam** From Convolutional Sparse Coding to Convolutional Neural Networks

More Information

- **Slides of the presentations**
  - http://vision.jhu.edu/tutorials/ICCV17-Tutorial-Math-Deep-Learning.htm

- **Paper,**
  - Conference on Decision and Control, December 2017

Mathematics of Deep Learning

René Vidal Joan Bruna Raja Giryes Stefano Soatto

Abstract— **Recently there has been a dramatic increase in the performance of recognition systems due to the introduction of deep architectures for representation learning and classification. However, the mathematical reasons for this success remain elusive. This tutorial will review recent work that aims to provide a mathematical justification for several properties of deep networks, such as global optimality, geometric stability, and invariance of the learned representations.**

sigmoidal activations are universal function approximators [5], [6], [7], [8]. However, the capacity of a wide and shallow network can be replicated by a deep network with significant improvements in performance. One possible explanation is that deeper architectures are able to better capture invariant properties of the data compared to their shallow counterparts. In computer vision, for example, the category of an object