Stat 212b: Topics in Deep Learning
Lecture 1: Introduction

Joan Bruna
UC Berkeley

Logistics

- Office Hours
  - Tuesdays 4-6pm, Evans 419
- Course evaluation
  - Paper Reviewing (30%): two papers during the semester
  - Final Project (70%): you can choose among
    - Oral Paper presentation
    - Tiny research project
    - Contribute to an open-source software package (Torch, Theano, Caffe)
  - Final Project Proposal due April 1st
- B-Courses (public access): https://bcourses.berkeley.edu/courses/1413088
- Github (public access): https://github.com/joanbruna/stat212b

What this course is NOT about

- Exhaustive review of state-of-the-art
  - Although we will talk about recent work
  - CS280 is focused on computer vision tasks.
  - CS188/287 develops some (deep) Reinforcement Learning.

What this course is NOT about

- Exhaustive review of state-of-the-art
  - Although we will talk about recent work
  - CS280 is focused on computer vision tasks.
  - CS188/287 develops some (deep) Reinforcement Learning.

- Hands-on implementations
  - Although (you!) will implement stuff

What this course is NOT about

- Exhaustive review of state-of-the-art
  - Although we will talk about recent work
  - CS280 is focused on computer vision tasks.
  - CS188/287 develops some (deep) Reinforcement Learning.

- Hands-on implementations
  - Although (you!) will implement stuff

- "Stratosferic" AI
  - Although we will talk a bit about Reasoning, Memory and sequence-to-sequence learning.

What this course is about

- Mathematical models of Deep Convolutional Networks
- Supervised and Unsupervised learning using Deep models.
- Applications to computer vision, speech and time series.
- Relationships between Deep Learning and "classic" models.
- Open mathematical/statistical questions.

Course Objectives

1. Good understanding of Convolutional (and recurrent) Networks

Course Objectives

1. Good understanding of Convolutional (and recurrent) Networks

2. Overview of current DL research

Course Objectives

1. Good understanding of Convolutional (and recurrent) Networks

2. Overview of current DL research

3. Identification of "good" open problems.

Course Objectives

1. Good understanding of Convolutional (and recurrent) Networks

2. Overview of current DL research

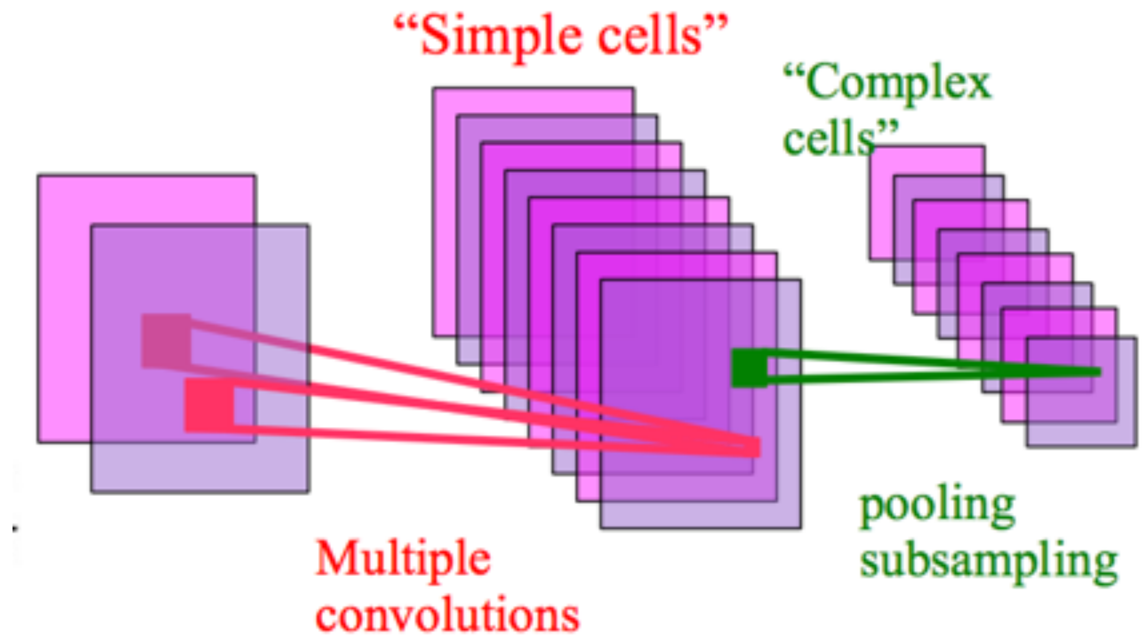
3. Identification of "good" open problems.

4. Feedback!

Deep Learning (take 1)

Early Hierarchical Feature Models for Vision

- Hubel & Wiesel [60s] Simple & Complex cells architecture:



- Fukushima’s Neocognitron [70s]

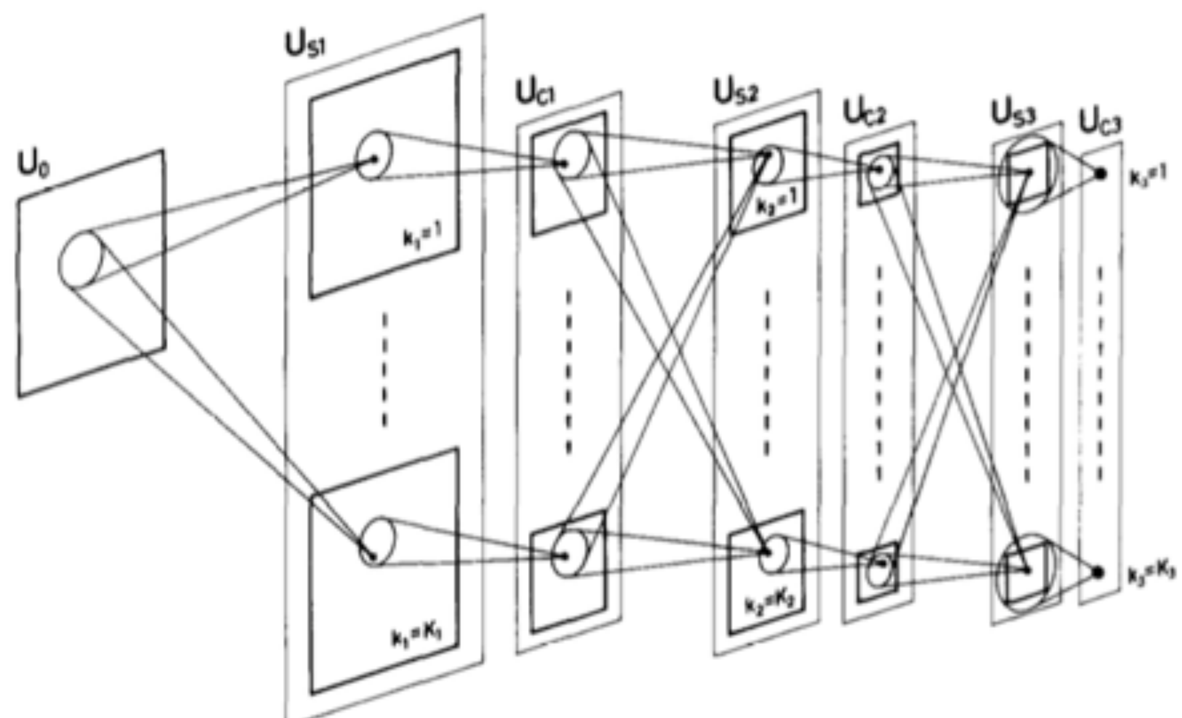
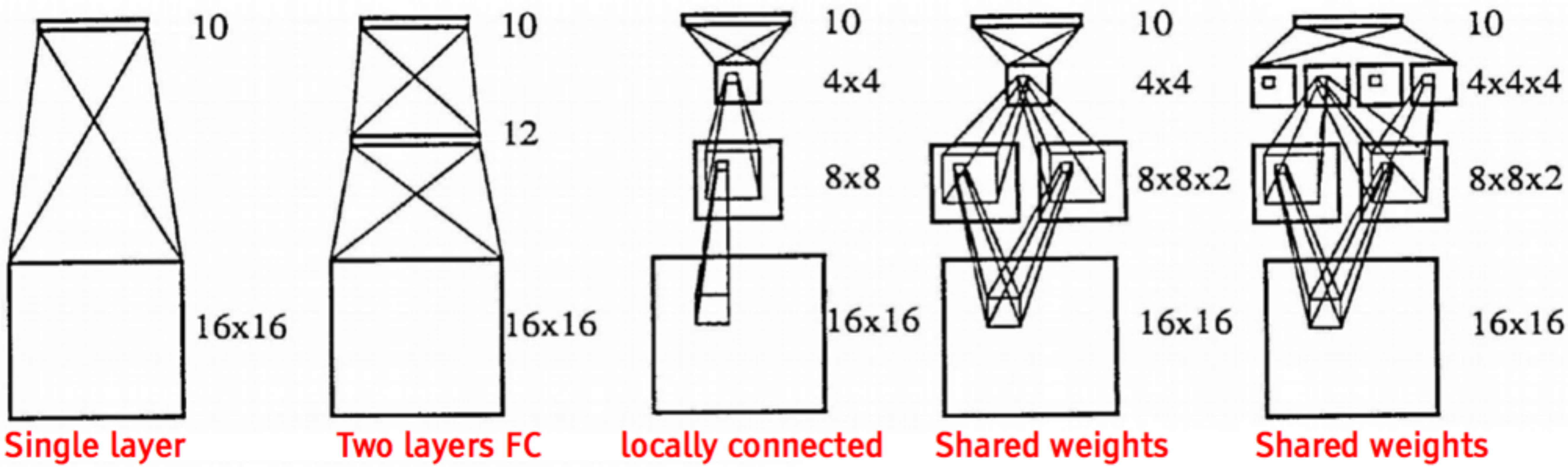


Fig. 2. Schematic diagram illustrating the interconnections between layers in the neocognitron

figures from Yann LeCun’s CVPR’15 plenary

Early Hierarchical Feature Models for Vision

- Yann LeCun's Early ConvNets [80s]:

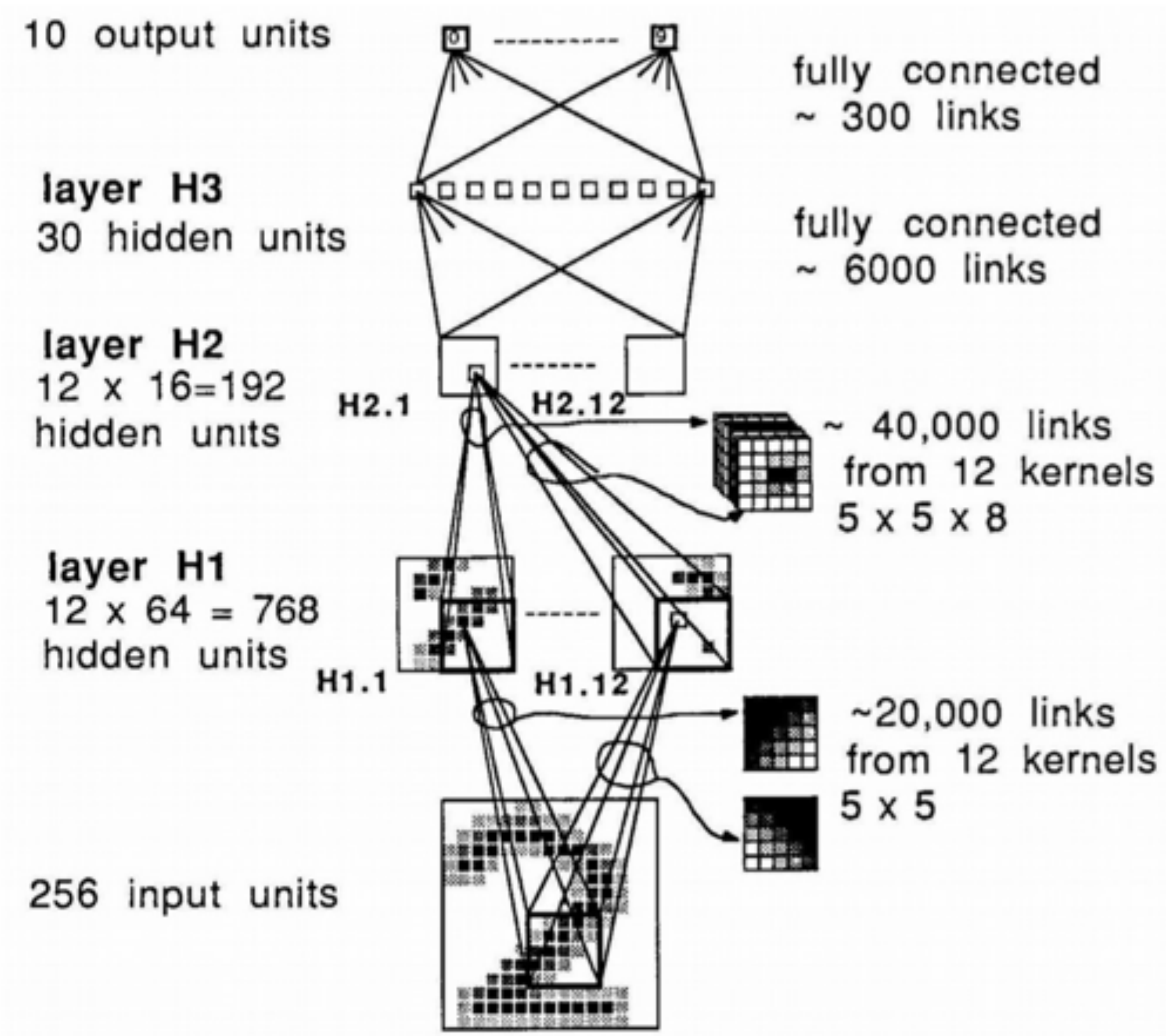


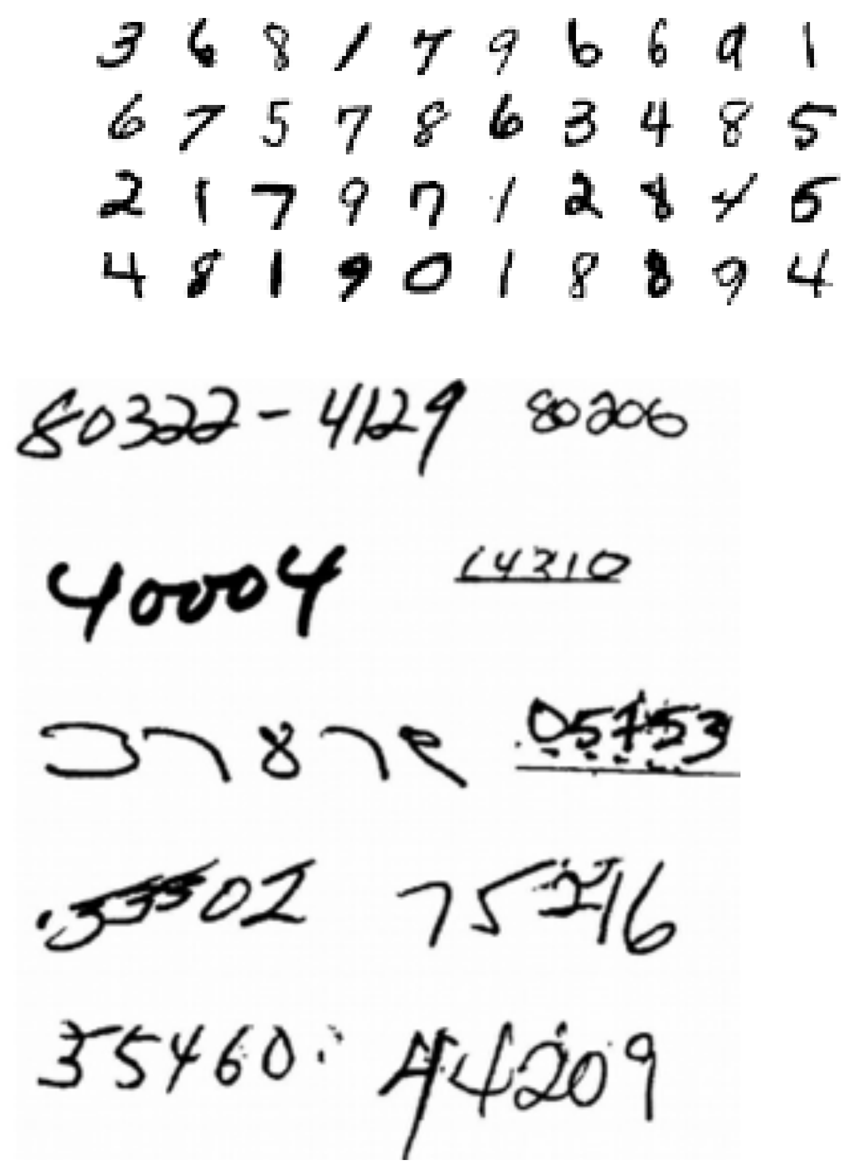
  - Used for character recognition
  - Trained with back propagation.

figures from Yann LeCun's CVPR'15 plenary

Deep Learning pre-2012

- Despite its very competitive performance, deep learning architectures were not widespread before 2012.
  - State-of-the-art in handwritten pattern recognition [LeCun et al. '89, Ciresan et al, '07, etc]





figures from Yann LeCun's CVPR'15 plenary

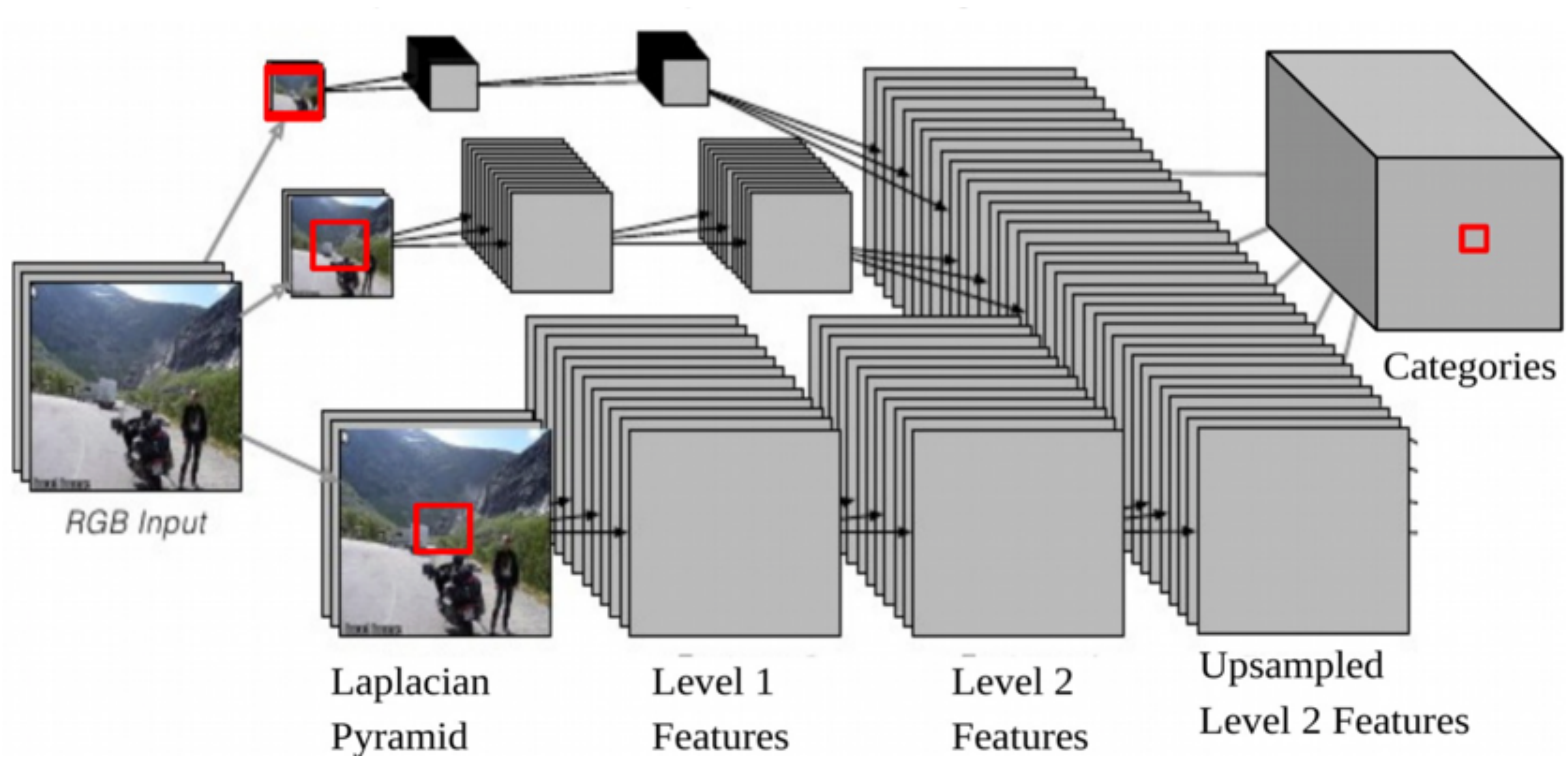
Deep Learning pre-2012

- Despite its very competitive performance, deep learning architectures were not widespread before 2012.
  - Face detection [Vaillant et al'93,'94 ; Osadchy et al, '03, '04, '07]

(Yann's Family)

Deep Learning pre-2012

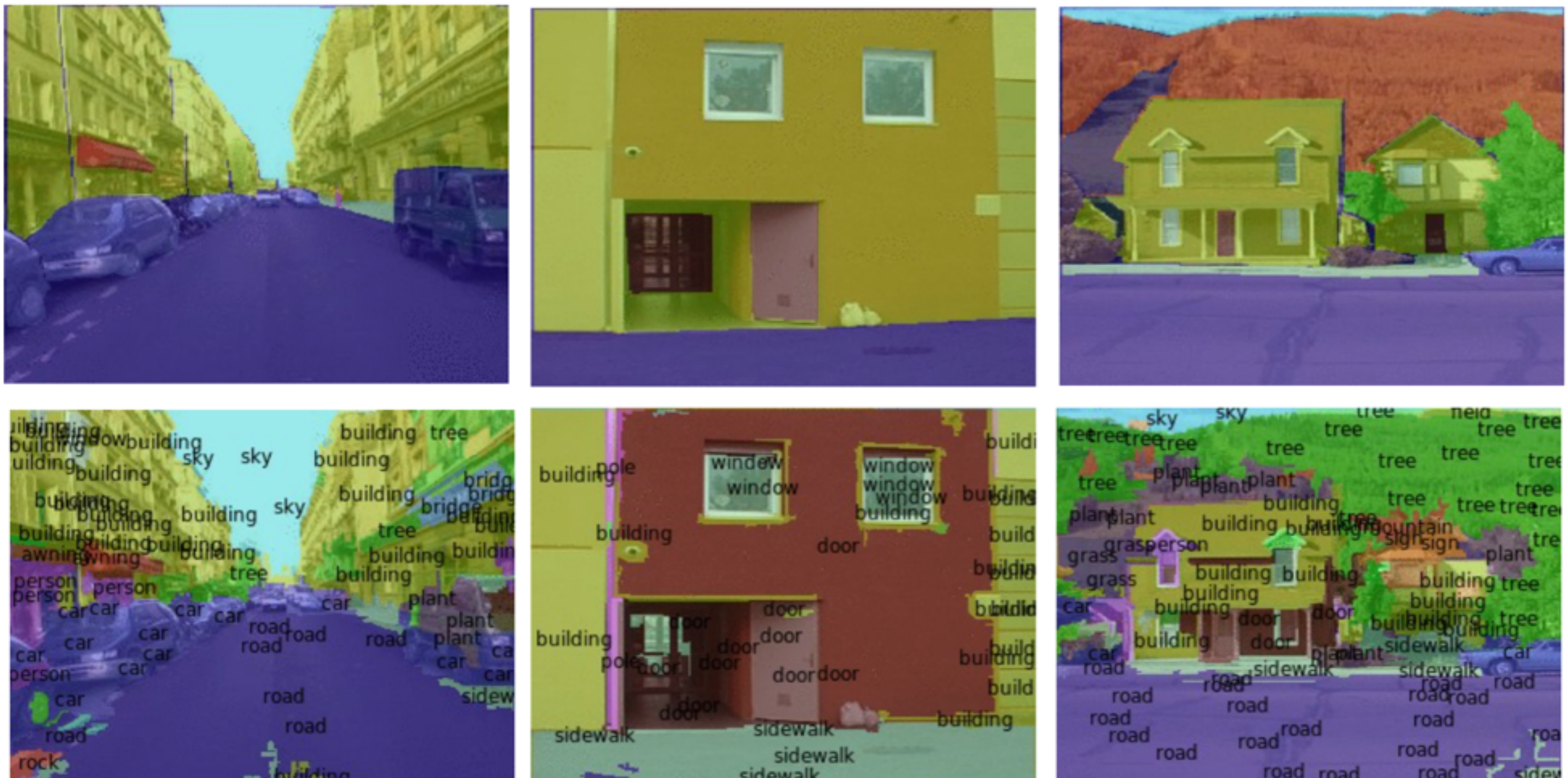
- Despite its very competitive performance, deep learning architectures were not widespread before 2012.
  - Scene Parsing [Farabet et al, '12,'13]



figures from Yann LeCun's CVPR'15 plenary

Deep Learning pre-2012

- Despite its very competitive performance, deep learning architectures were not widespread before 2012.
  - Scene Parsing [Farabet et al, '12,'13]



figures from Yann LeCun's CVPR'15 plenary

Long Story Short

- "A class of parametrized non-linear representations encoding appropriate domain knowledge (invariance and stationarity) that can be (massively) optimized efficiently using stochastic gradient descent"

Long Story Short

- "A class of parametrized non-linear representations encoding appropriate domain knowledge (invariance and stationarity) that can be (massively) optimized efficiently using stochastic gradient descent"

$$\Phi(x\,,\,\Theta) = \rho(W_L(\rho(W_{L-1}\dots\rho(W_1(x))\dots)$$

$$W_i \;:\; \text{Convolutional Tensors}$$

$$\rho(\cdot) \;:\; \text{point-wise thresholding}$$

Long Story Short

- "A class of parametrized non-linear representations encoding appropriate domain knowledge (invariance and stationarity) that can be (massively) optimized efficiently using stochastic gradient descent"

$$\Phi(x\,,\,\Theta) = \rho(W_L(\rho(W_{L-1} \ldots \rho(W_1(x)) \ldots)$$

$$W_i \ : \ \text{Convolutional Tensors}$$

$$\rho(\cdot) \ : \ \text{point-wise thresholding}$$

Given labeled data $\{x_i, y_i\}_i$, solve using online stochastic optimization:

$$\hat{y}_i(\Theta) = \text{softmax}(\overline{\Phi}(x_i, \Theta))$$

$$\Theta^* \leftarrow \arg\min_{\Theta} \ E(\Theta) = \sum_i \ell(\hat{y}_i(\Theta), y_i)$$

Long Story Short

– Too many parameters to learn from few labeled examples.

– "I know my features are better for this task".

– Non-convex optimization? No, thanks.

– Black-box model, no interpretability.

– Mostly point estimates: Non inferential.

Deep Learning (take 2)

Deep Learning Golden age in Vision

- 2012-2014 Imagenet results:

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

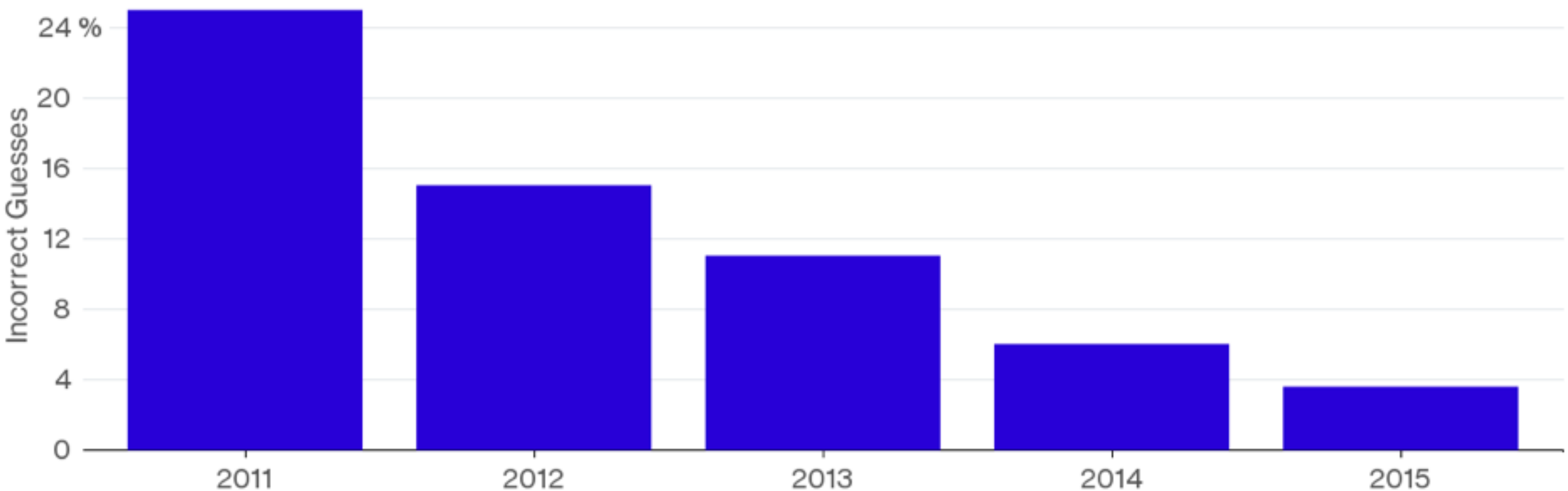
- 2015 results: MSRA under **3.5%** error. (using a CNN with 150 layers!)

figures from Yann LeCun's CVPR'15 plenary

Progress in large-scale Image Classification

Computers Stop Squinting and Open Their Eyes

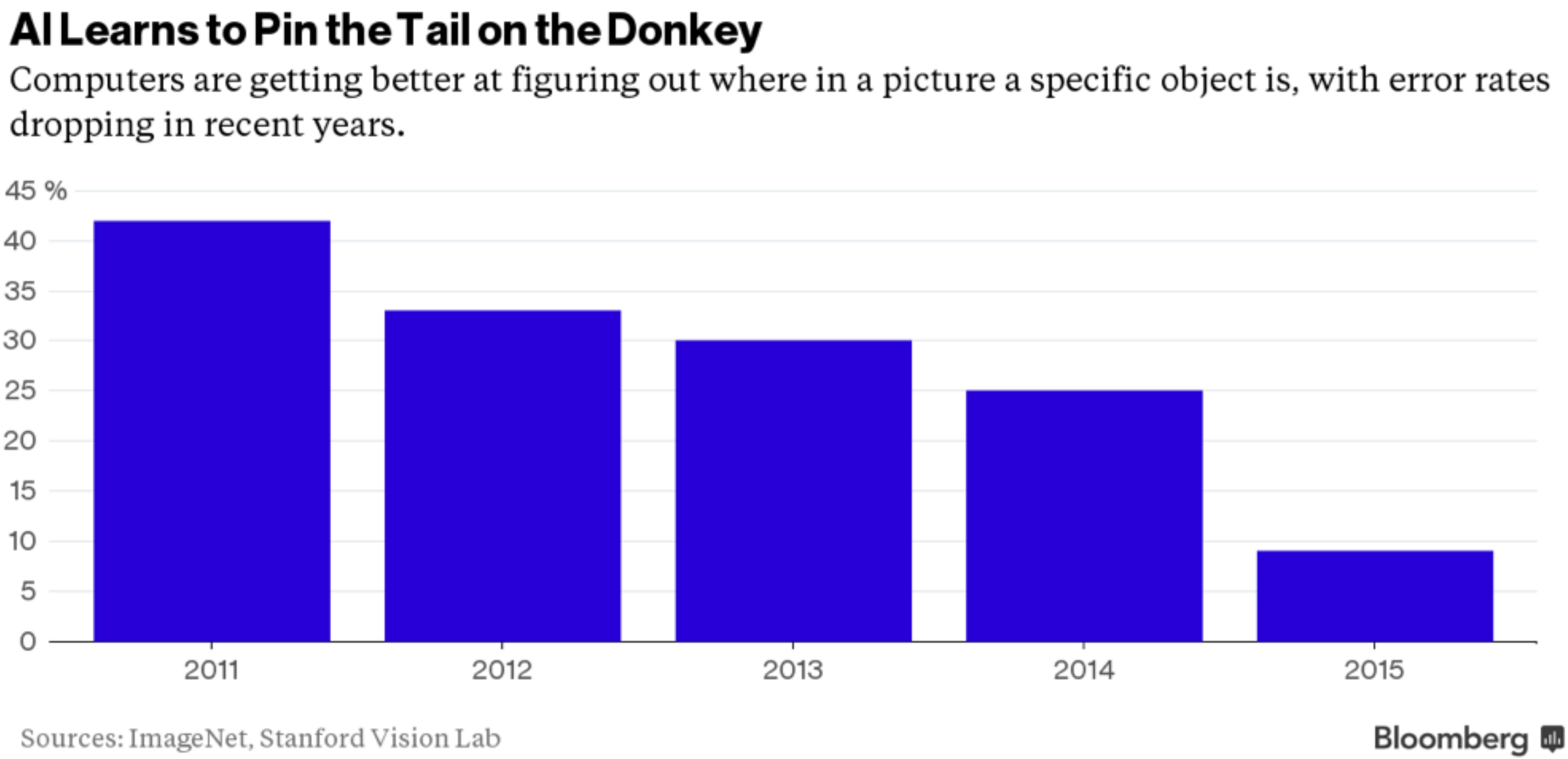
Error rates on a popular image recognition challenge have fallen dramatically since the advent of deep learning systems in the 2012 competition.



Sources: ImageNet, Stanford Vision Lab

Bloomberg

Progress in Object Localization



Some Puzzling Questions

- What made this result possible?

Some Puzzling Questions

- What made this result possible?
  - Larger training sets (1.2 million, high-resolution training samples, 1000 object categories)
  - Better Hardware (GPU)
  - Better Learning Regularization (eg Dropout)
  - Better Optimization Conditioning (eg Batch Normalization)

Some Puzzling Questions

- What made this result possible?
  - Larger training sets (1.2 million, high-resolution training samples, 1000 object categories)
  - Better Hardware (GPU)
  - Better Learning Regularization (eg Dropout)
  - Better Optimization Conditioning (eg Batch Normalization)

- Is this just for a particular dataset?

- Is this just for a particular task?

Some Puzzling Questions

- What made this result possible?
  - Larger training sets (1.2 million, high-resolution training samples, 1000 object categories)
  - Better Hardware (GPU)
  - Better Learning Regularization (eg Dropout)
  - Better Optimization Conditioning (eg Batch Normalization)

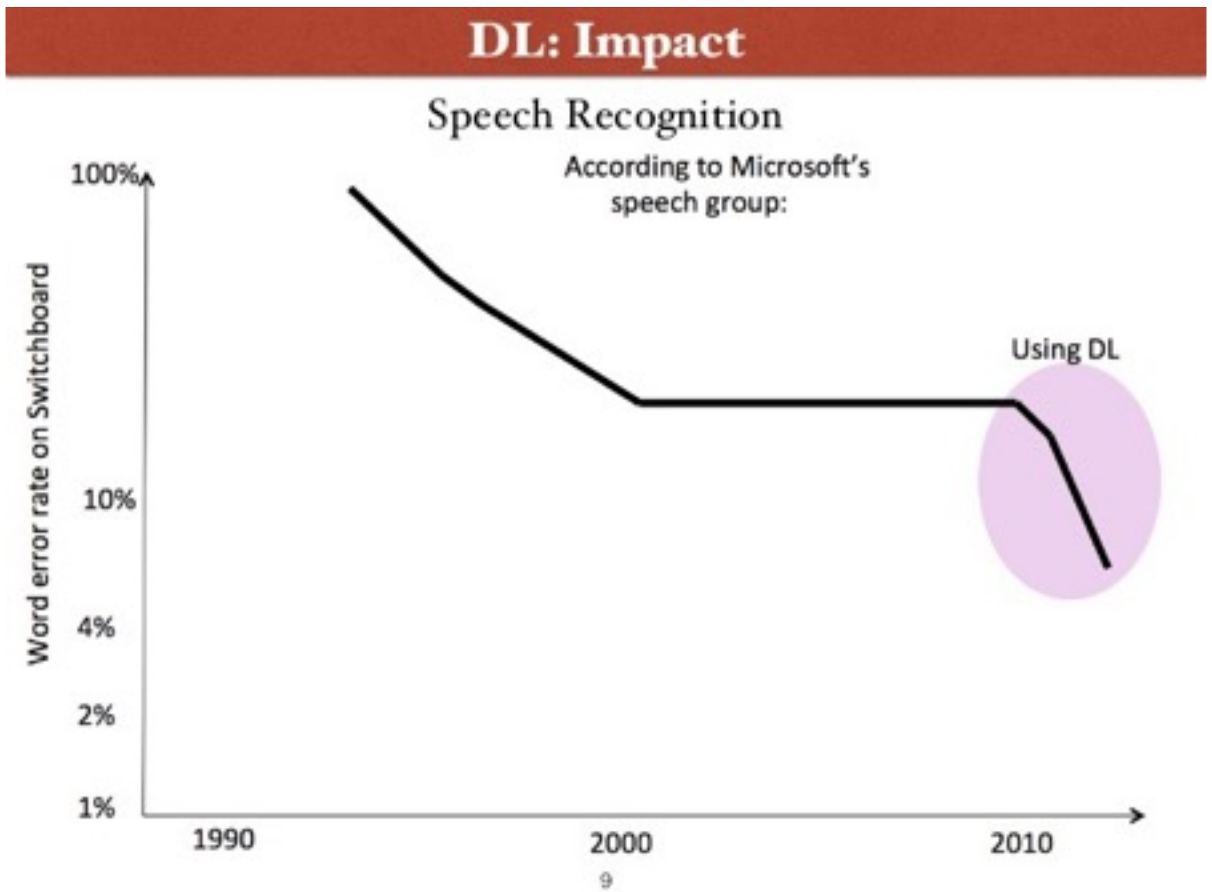
- Is this just for a particular dataset?

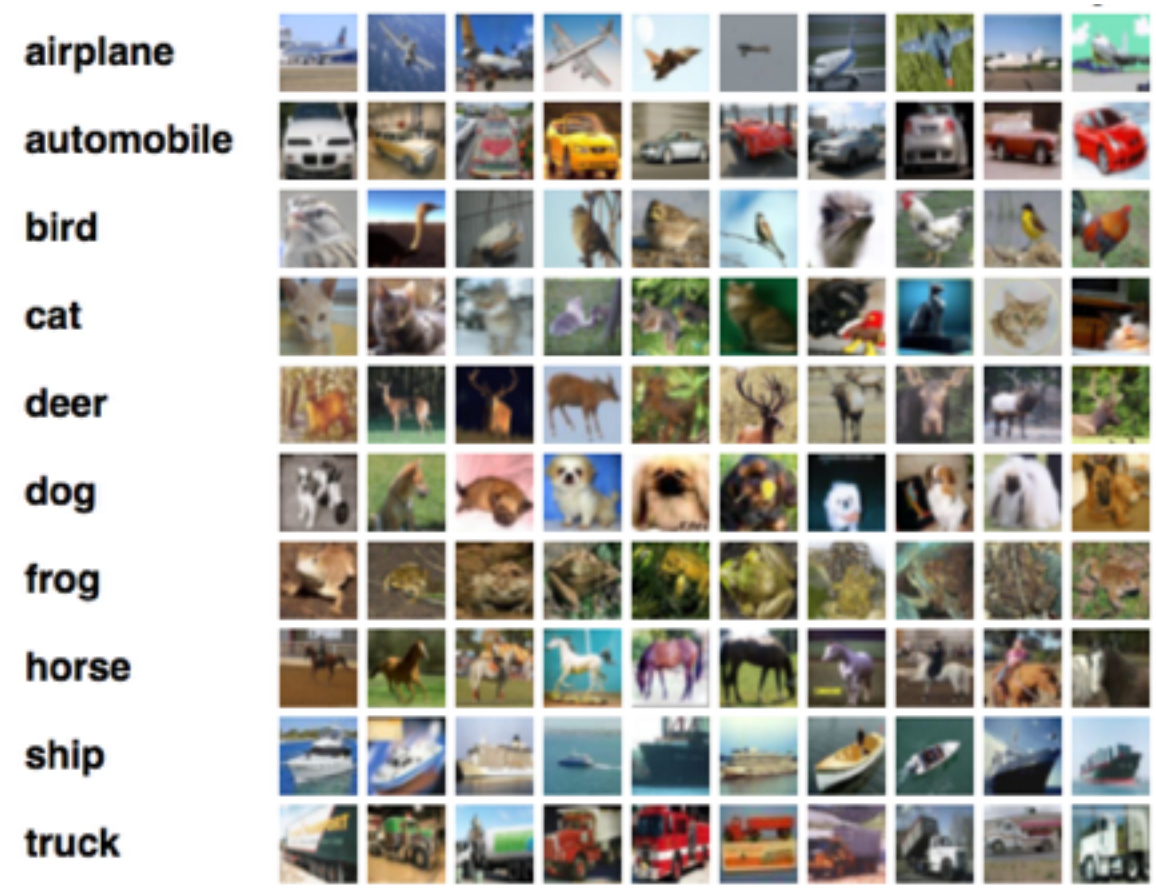
- Is this just for a particular task?

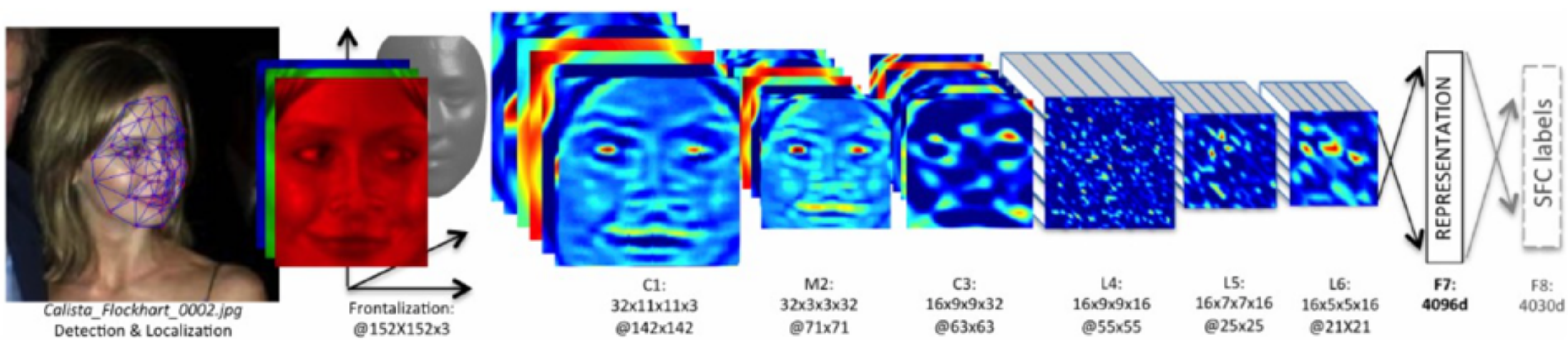
- Why are these architectures so efficient?
  - *We'll look for mathematical and statistical reasons.*

Is it just for a particular dataset?

- No. Nowadays CNNs hold the state-of-the-art on virtually any object classification task.







figures from Yann LeCun's NIPS'15 tutorial

Is it just for a particular dataset?

- No. CNNs work well on domains where there is a low-dimensional geometric structure.

AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery

Izhar Wallach
Atomwise, Inc.
izhar@atomwise.com

Michael Dzamba
Atomwise, Inc.
misko@atomwise.com

Abraham Heifets
Atomwise, Inc.
abe@atomwise.com

Abstract

Deep convolutional neural networks comprise a subclass of deep neural networks (DNN) with a constrained architecture that leverages the spatial and temporal structure of the domain they model. Convolutional networks achieve the best predictive performance in areas such as speech and image recognition by hierarchically composing simple local features into complex models. Although DNNs have been used in drug discovery for QSAR and ligand-based bioactivity predictions, none of these models have benefited from this powerful convolutional architecture. This paper introduces AtomNet, the first structure-based, deep convolutional neural network designed to predict the bioactivity of small molecules for drug dis-

Is it just for a particular task?

- No. CNN architectures also obtain state-of-the-art performance on many other tasks:

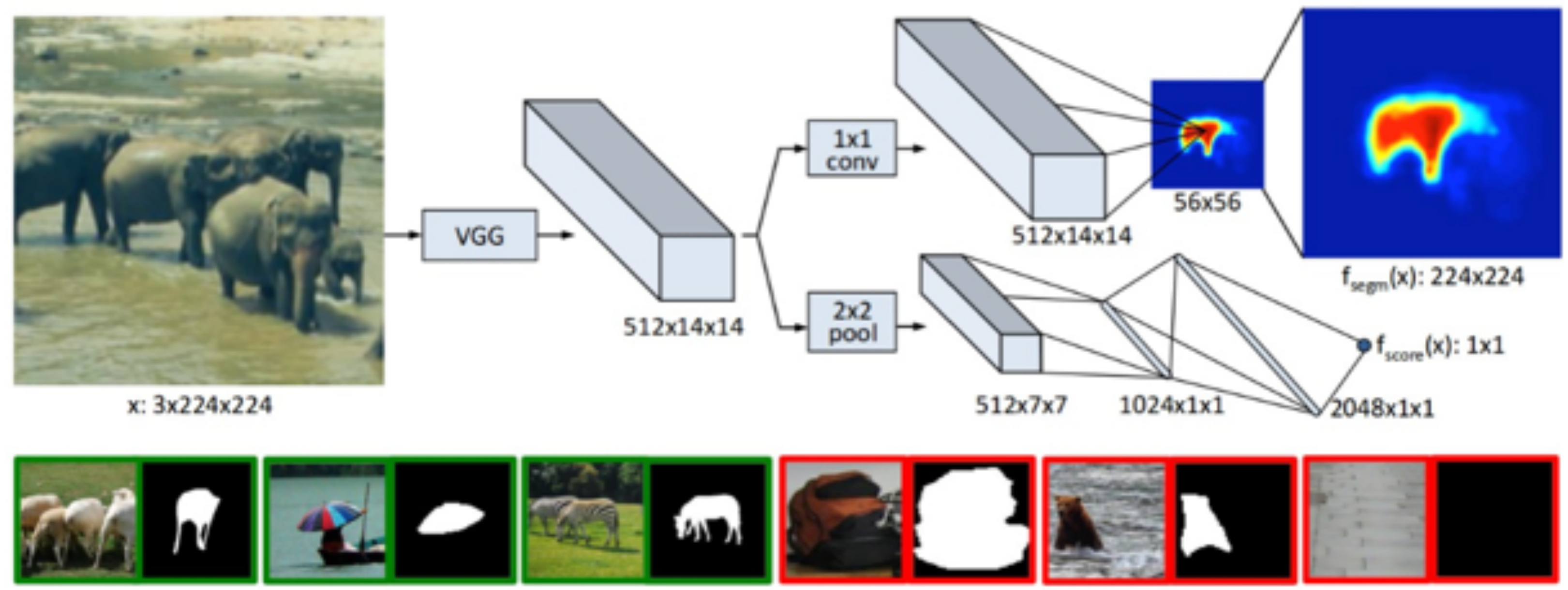
Object Localization
[R-CNN, HyperColumns, Overfeat, etc.]

Pose estimation [Tomson et al, CVPR'15]

figures from Yann LeCun's CVPR'15 plenary

Is it just for a particular task?

- No. CNN architectures also obtain state-of-the-art performance on other tasks:



- Segmentation [Pinhero, Collobert, Dollar, ICCV'15]

figures from Yann LeCun's CVPR'15 plenary

Is it just for a particular task?

- No. CNN architectures also obtain state-of-the-art performance on other tasks:
    - Image Captioning [Vinyals et al'14, Karpathy et al '14, Donahue et al'14, Kiros et al'14,MSR'14]

A person riding a motorcycle on a dirt road.

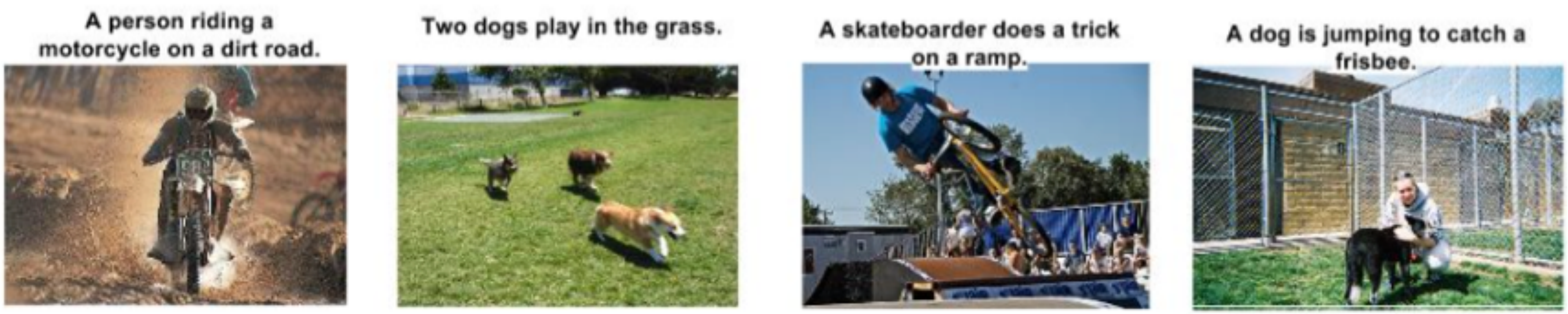
Two dogs play in the grass.

A skateboarder does a trick on a ramp.

A dog is jumping to catch a frisbee.

Is it just for a particular task?

- No. CNN architectures also obtain state-of-the-art performance on other tasks:
  - Image Captioning [Vinyals et al'14, Karpathy et al '14, Donahue et al'14, Kiros et al'14,MSR'14]



  - Image Super-Resolution [MSR'14]

  - Optical Flow estimation [Zontar & LeCun, '15]
  - etc..

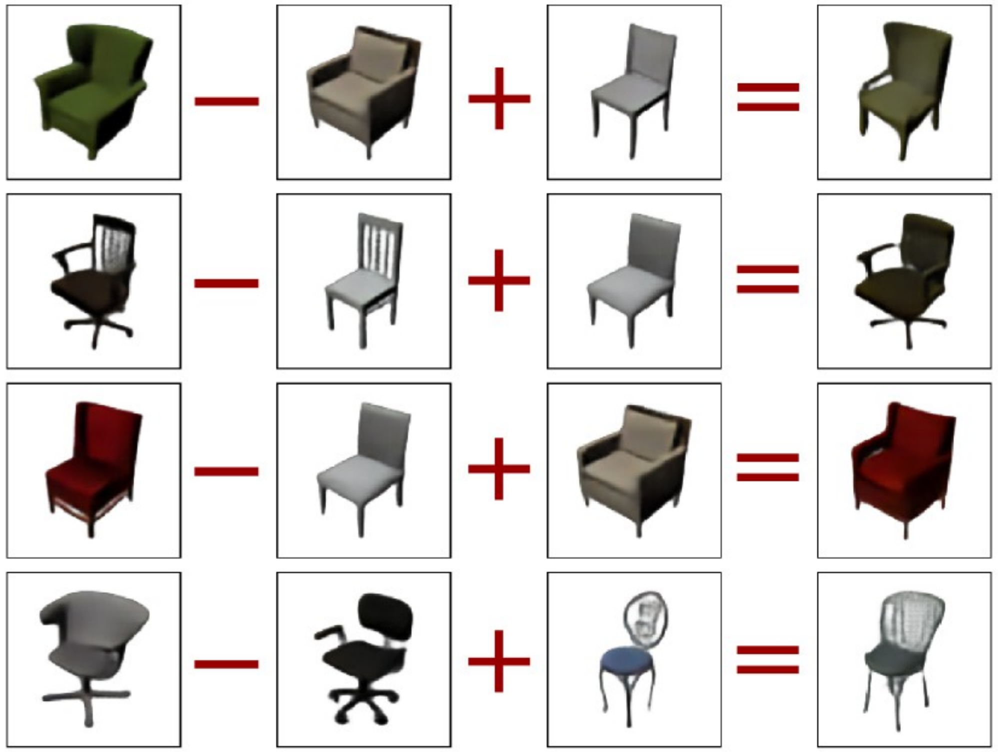
Beyond Supervised Learning

Beyond Supervised Learning

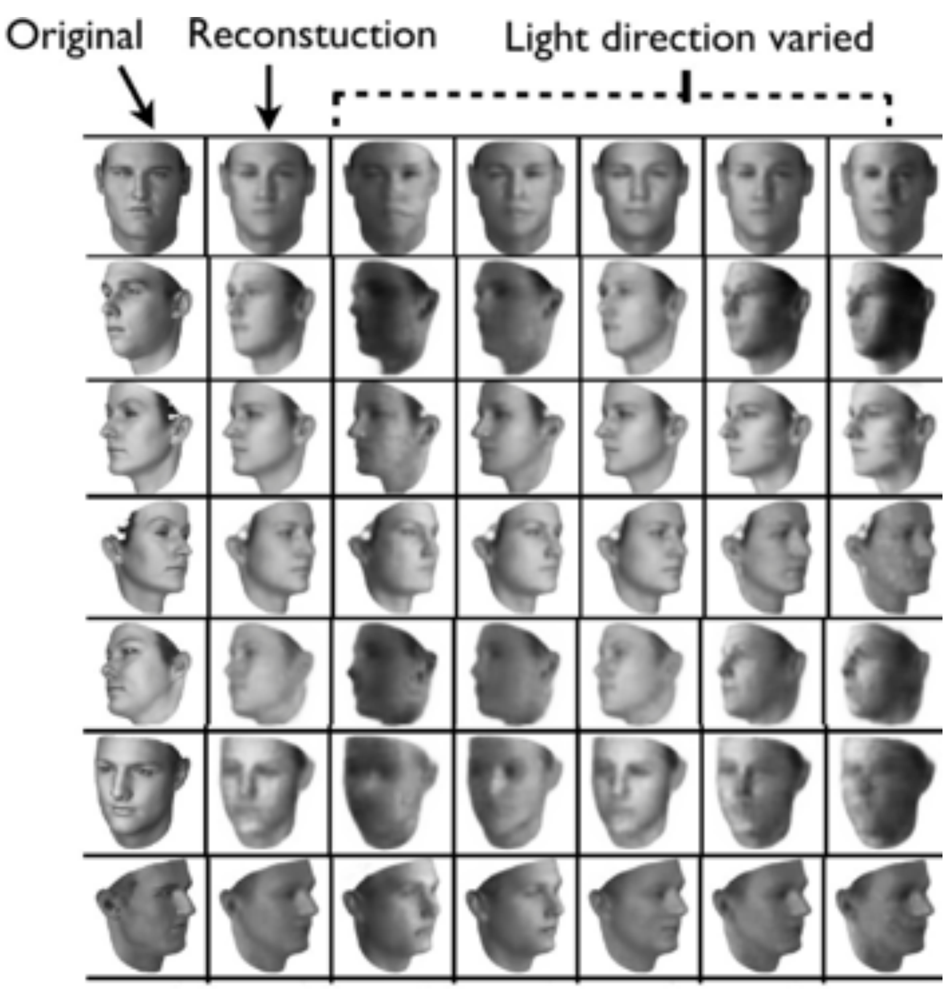
- Deep Mind success (2013-now)

Beyond Supervised Learning

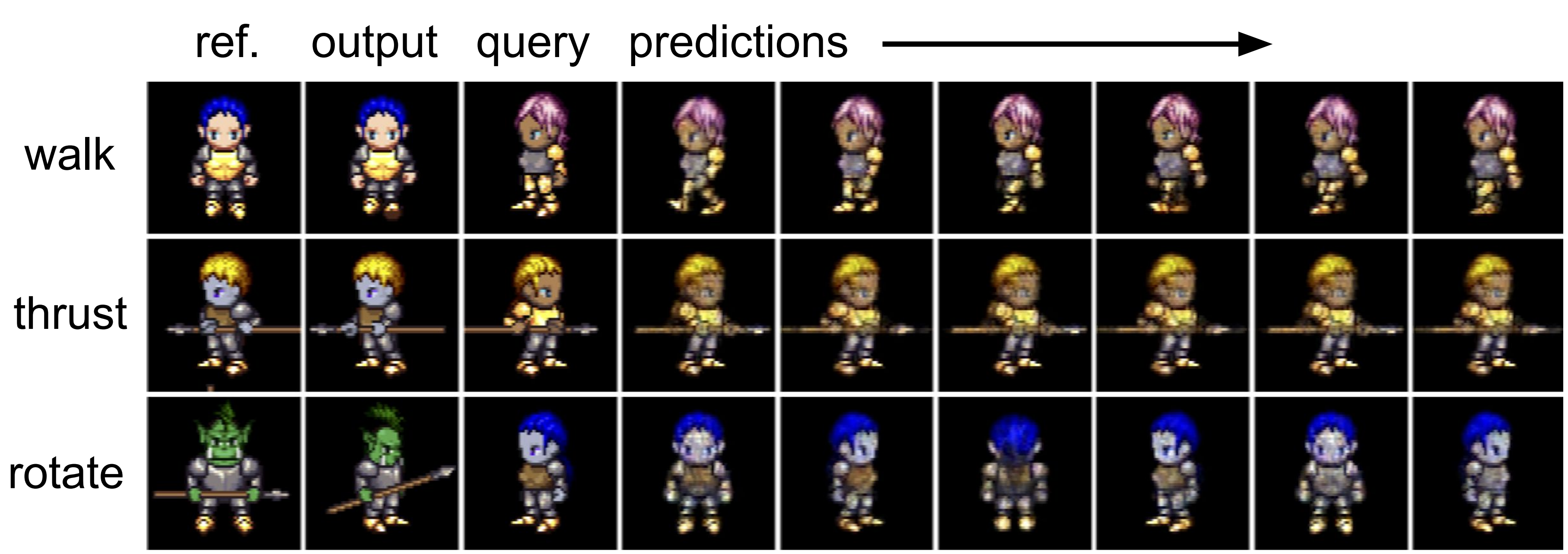
- Visual analogies using CNNs:



[Dosovitsky et al '14]



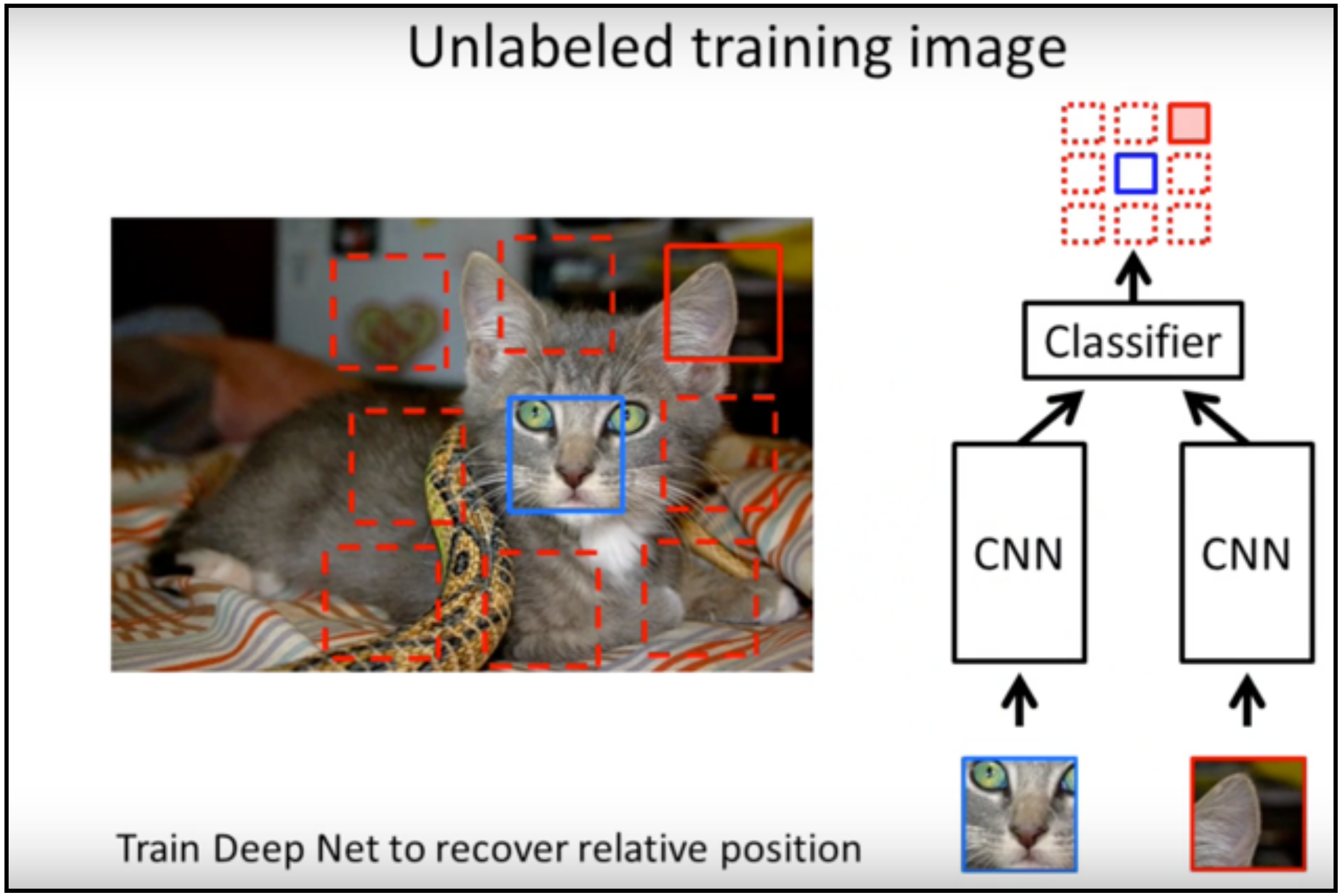
[Kulkarni et al'15]



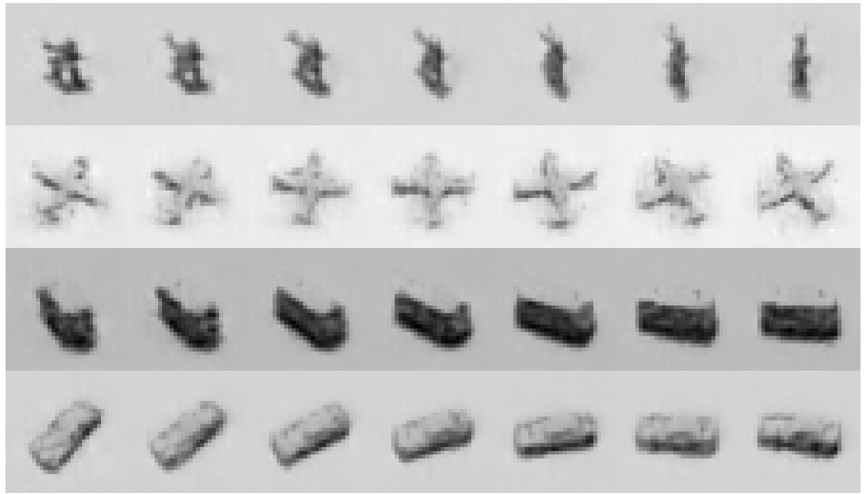
[Reed et al'15]

From Supervised to "Self-Supervised" Learning

- Exploit spatio-temporal structure to constrain good image representations, eg:



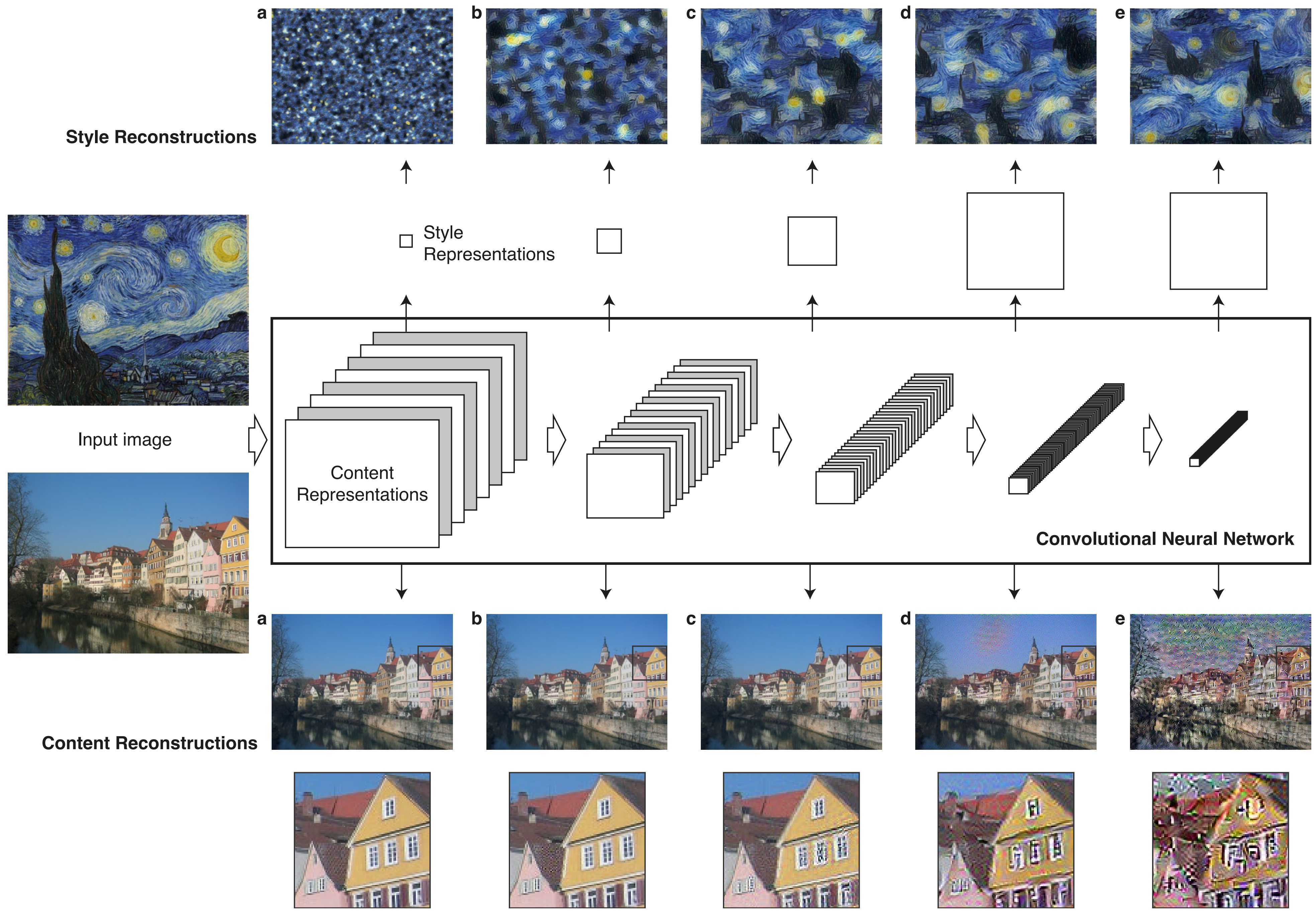
[Doersch et al'15]



[Goroshin et al'15]

Texture and Geometry

- CNN Representations arising from large-scale classification "disentangle" texture from geometry:



[Gathys et al'15]

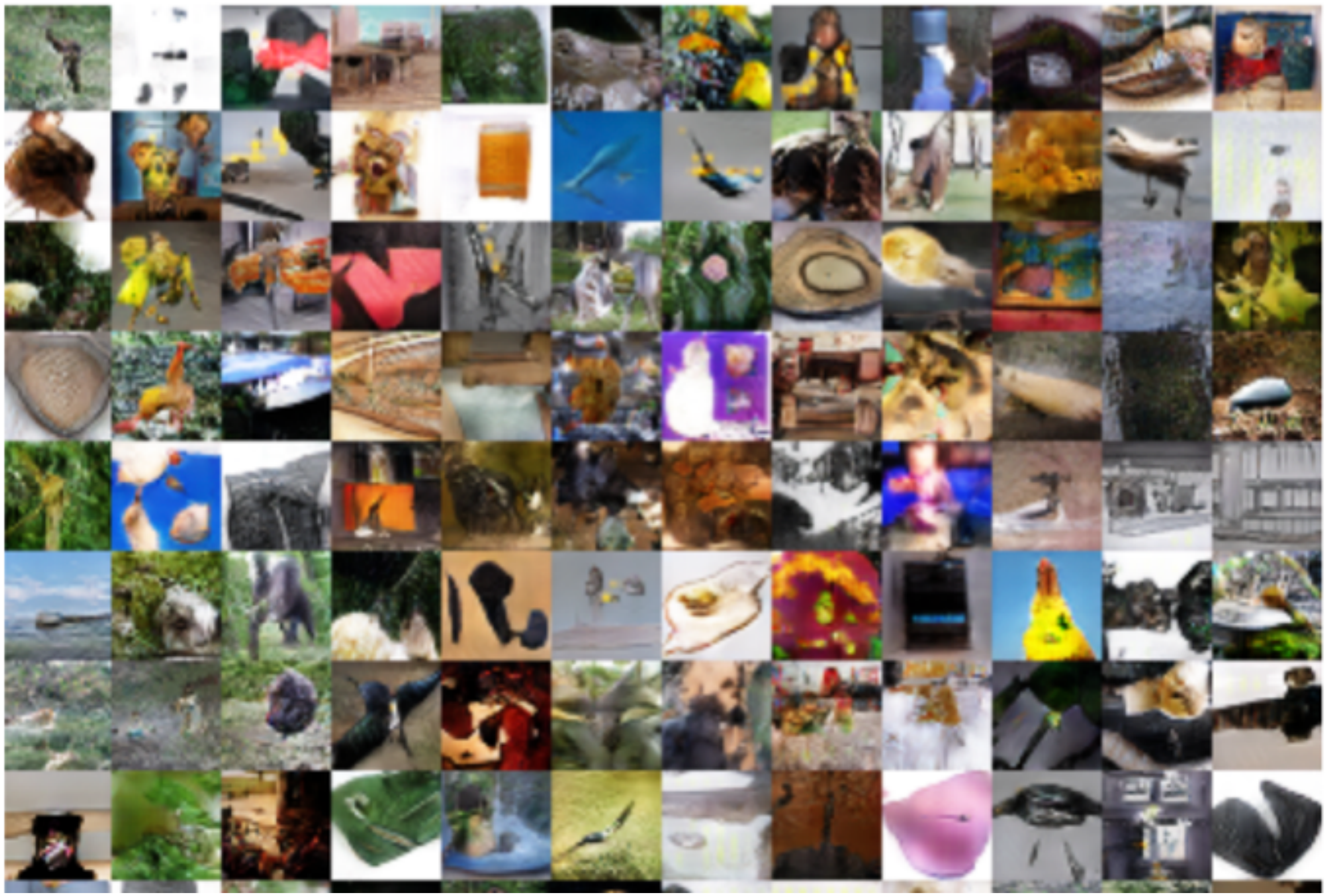
Texture and Geometry

- CNN Representations arising from large-scale classification “disentangle” texture from geometry:

[Gathys et al’15]

Generative Models of Natural Images

- CNNs can also be used to generate images, using appropriate loss and optimization 'tricks':



DC-GAN [Radford, Metz & Chintala,'15]

Generative Models of Natural Images

- CNNs can also be used to generate images, using appropriate loss and optimization 'tricks':

  - DC-GAN [Radford, Metz & Chintala,'15]

- Convolutional Deep Learning models thus appear to capture high level image properties more efficiently than previous models.

- Convolutional Deep Learning models thus appear to capture high level image properties more efficiently than previous models.
  - Highly Expressive Representations capturing complex geometrical and statistical patterns.
  - Excellent generalization: "beating" the curse of dimensionality
  - The representation extracts geometry and texture "automatically"

- Convolutional Deep Learning models thus appear to capture high level image properties more efficiently than previous models.

- Which architectural choices might explain this advantage mathematically?

- Convolutional Deep Learning models thus appear to capture high level image properties more efficiently than previous models.

- Which architectural choices might explain this advantage mathematically?
  - Role of non-linearities?
  - Role of convolutions?
  - Role of depth?
  - Interplay with geometrical, class-specific invariants?

- Convolutional Deep Learning models thus appear to capture high level image properties more efficiently than previous models.

- Which architectural choices might explain this advantage mathematically?

- Which optimization choices might explain this advantage?

- Convolutional Deep Learning models thus appear to capture high level image properties more efficiently than previous models.

- Which architectural choices might explain this advantage mathematically?

- Which optimization choices might explain this advantage?
  - Presence of local minima or saddle points?
  - Equivalence of local solutions?
  - Role of Stochastic optimization?
  - Role of Normalization?

Sequence learning with RNNs

- Images and Sounds are subject to the physical world (and to harmonic analysis).
- Other important data of interest is not: language, robotics.

Sequence learning with RNNs

- Images and Sounds are subject to the physical world (and to harmonic analysis).
- Other important data of interest is not: language, robotics.

- Generic setup:

$$S = (s_0, s_1, \ldots, s_k, \ldots) \ , \ s_k \in \mathcal{X}$$

- Sequence modeling:

$$p(S) = p(s_0) \prod_k p(s_k \mid s_0 \ldots s_{k-1})$$

- Sequence translation:

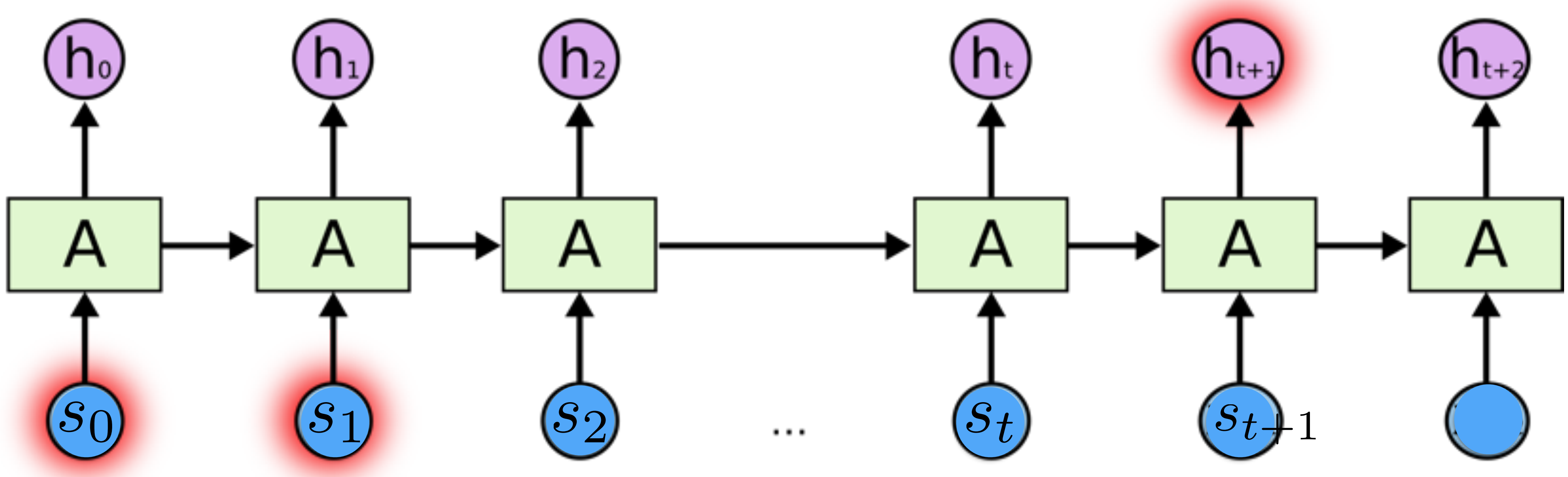
$$p(S \mid R) = p(s_0 \mid R) \prod_k p(s_k \mid s_0 \ldots s_{k-1}, R)$$

Sequence learning with RNNs

- Curse of dimensionality is broken by projecting the past information into a finite-dimensional space and using recurrence:

$$p(s_k \mid s_0 \ldots, s_{k-1}) = f(s_k, h_k)$$

$$h_k = g(h_{k-1}, s_k) \ , \ h_k \in \mathbb{R}^p \ .$$

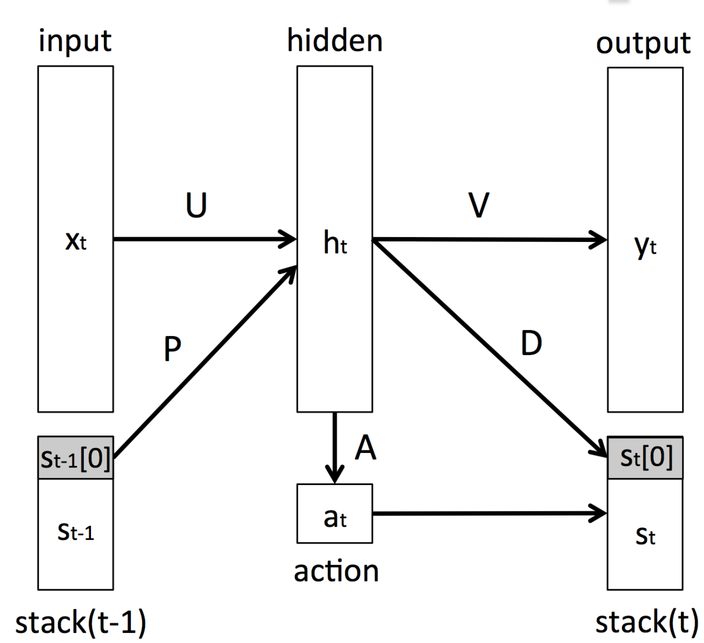


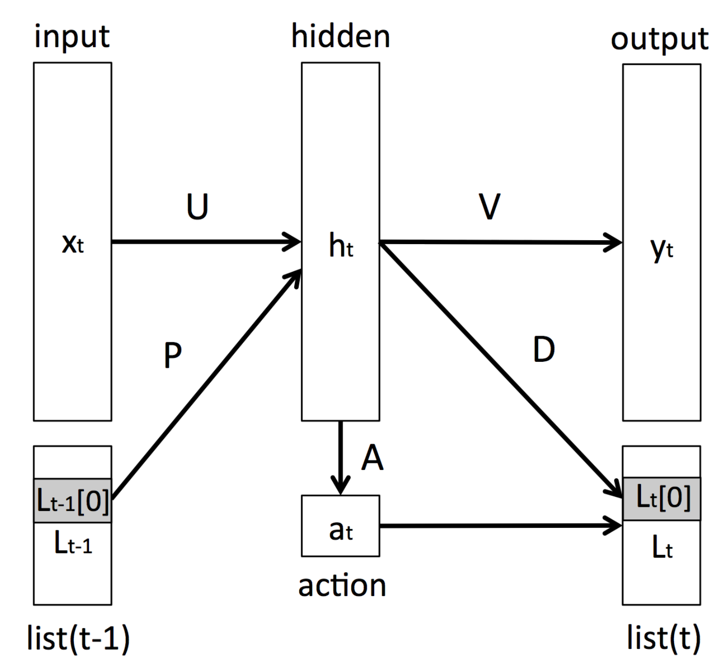
[from Chris Olah's blog]

Recent Sequential models

- Attention mechanisms [Badhanu et al'14, DeepMind'14-15, …]
  - Introduce a controlled form of non-stationarity in the model

- *Differentiable Memory* structures:
  - LSTM [Hochreiter & Schmidhuber]
  - Tapes [NTM, Graves et al'14]
  - Arrays [Memory Nets, Weston et al'14]
  - Stacks [Joulin & Mikolov'15]
  - …





[Joulin & Mikolov,'15]

Applications of Recurrent Models

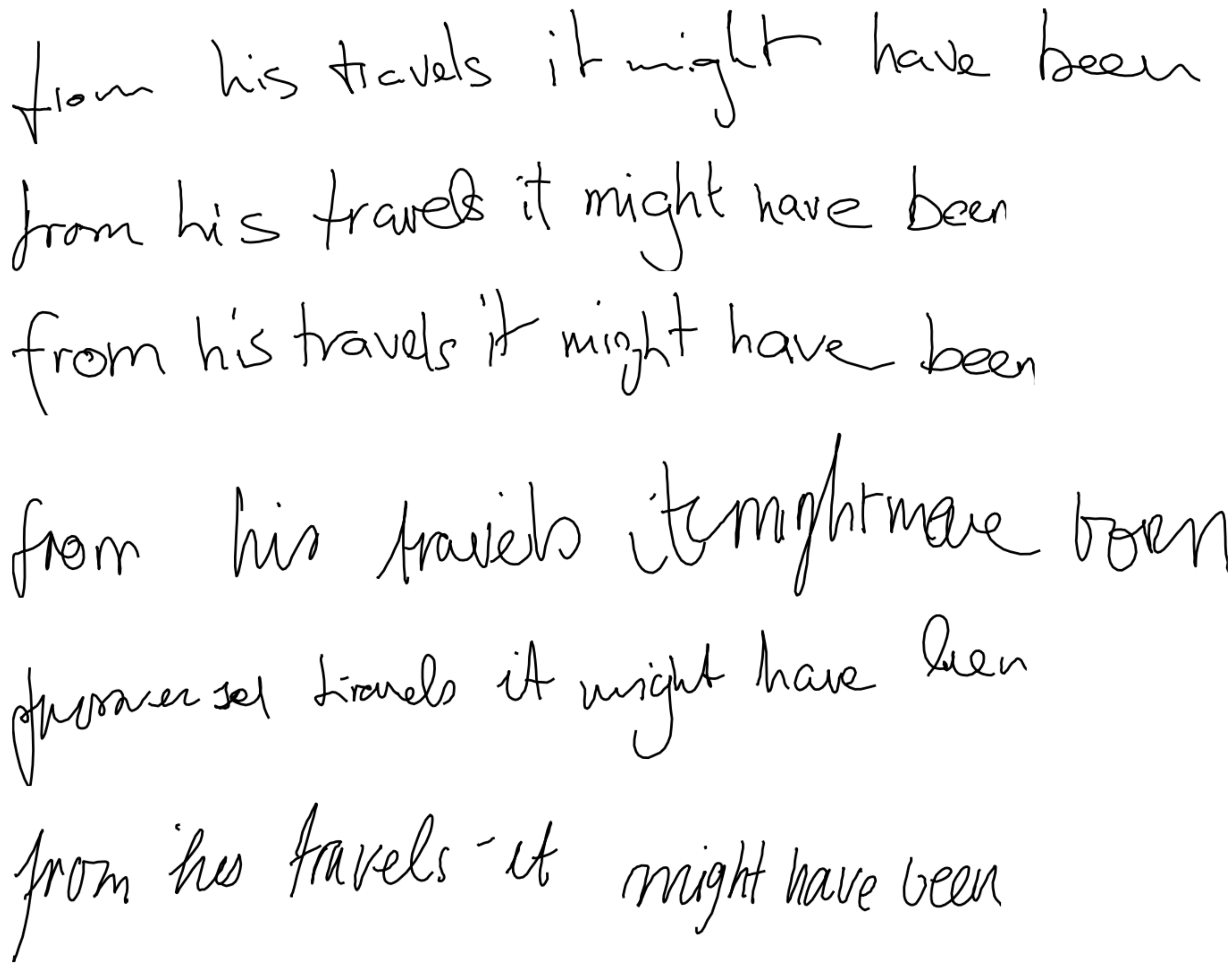
- Language modeling (see Andrej Karpathy's blog)
- Machine Translation:

Source	An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.
Reference	Le privilège d'admission est le droit d'un médecin, en vertu de son statut de membre soignant d'un hôpital, d'admettre un patient dans un hôpital ou un centre médical afin d'y délivrer un diagnostic ou un traitement.
RNNenc-50	Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.
RNNsearch-50	Un privilège d'admission est le droit d'un médecin d'admettre un patient à un hôpital ou un centre médical pour effectuer un diagnostic ou une procédure, selon son statut de travailleur des soins de santé à l'hôpital.
Google Translate	Un privilège admettre est le droit d'un médecin d'admettre un patient dans un hôpital ou un centre médical pour effectuer un diagnostic ou une procédure, fondée sur sa situation en tant que travailleur de soins de santé dans un hôpital.

[Badhanu et al]

Applications of Recurrent Models

- Synthesis models:



[A. Graves]

- Nonlinear Recurrent models can capture stationary information beyond second-order structure.
  - Comparisons with n-gram and convolutional models?
  - Extension to high-dimensional spaces?

- Attention and external memory models provide "non-stationary relief"
  - Role of memory layout?
  - Relationship to non-parametric models (eg K-Nearest Neighbors)

Deep Learning Approximation Theory

- Deep Networks define a class of "universal approximators":

Theorem [C'89, H'91] Let $\rho()$ be a bounded, non-constant continuous function. Let I_m denote the m-dimensional hypercube, and $C(I_m)$ denote the space of continuous functions on I_m. Given any $f \in C(I_m)$ and $\epsilon > 0$, there exists $N > 0$ and v_i, w_i, b_i, $i = 1 \ldots, N$ such that

$$F(x) = \sum_{i \leq N} v_i \rho(w_i^T x + b_i) \text{ satisfies}$$

$$\sup_{x \in I_m} |f(x) - F(x)| < \epsilon \ .$$

Deep Learning Approximation Theory

- Deep Networks define a class of "universal approximators":

Theorem [C'89, H'91] Let $\rho()$ be a bounded, non-constant continuous function. Let I_m denote the m-dimensional hypercube, and $C(I_m)$ denote the space of continuous functions on I_m. Given any $f \in C(I_m)$ and $\epsilon > 0$, there exists $N > 0$ and v_i, w_i, b_i, $i = 1 \ldots, N$ such that

$$F(x) = \sum_{i \leq N} v_i \rho(w_i^T x + b_i) \quad \text{satisfies}$$

$$\sup_{x \in I_m} |f(x) - F(x)| < \epsilon \ .$$

- It guarantees that even a single hidden-layer network can represent any classification problem in which the boundary is locally linear (smooth).
- It does not inform us about which architectures are good…
- …Or how they relate to the optimization.

Deep Learning Estimation Theory

Theorem [Barron'92] The mean integrated square error between the estimated network $\hat{F}$ and the target function f is bounded by

$$O\left(\frac{C_f^2}{N}\right) + O\left(\frac{Nm}{K}\log K\right) ,$$

where K is the number of training points, N is the number of neurons, m is the input dimension, and C_f measures the global smoothness of f.

Deep Learning Estimation Theory

Theorem [Barron'92] The mean integrated square error between the estimated network $\hat{F}$ and the target function f is bounded by

$$O\left(\frac{C_f^2}{N}\right) + O\left(\frac{Nm}{K}\log K\right) ,$$

where K is the number of training points, N is the number of neurons, m is the input dimension, and C_f measures the global smoothness of f.

- Combines approximation and estimation error.
- Does not explain why online/stochastic optimization works better than batch normalization.
- Does not relate generalization error with choice of architecture.

Statistical Learning Theory

- One can compute the complexity or capacity of Neural Network models by measuring how many configurations can be shattered (VC dimension) [P. Bartlett et al, "Vapnik-Chervonenkis Dimension of Neural Nets"]

- The capacity of the network, if measured by the number of pieces in a piecewise linear approximation, increases exponentially with depth [Montufar, Pascanu et al, '14]

- These results quantify an upper bound on the empirical risk of deep neural networks
- They do not explain the superior generalization properties of CNNs versus models with similar capacity
- The bounds might be very pessimistic.

- Generally speaking, given high-dimensional data $X = (x_1, \ldots, x_n)$, we want to estimate a low-dimensional model characterizing the population.

Unsupervised Learning with Deep Networks

- Generally speaking, given high-dimensional data $X = (x_1, \ldots, x_n)$, we want to estimate a low-dimensional model characterizing the population.
- Why do we care?
  - Simulation environments, prediction, inverse problems, transfer learning.

Unsupervised Learning with Deep Networks

- Generally speaking, given high-dimensional data $X = (x_1, \ldots, x_n)$, we want to estimate a low-dimensional model characterizing the population.
- Why do we care?
  - Simulation environments, prediction, inverse problems, transfer learning.
- **Problem**: X is itself high-dimensional! (unless you believe in the low-dimensionality manifold hypothesis).

Unsupervised Learning with Deep Networks

- Generally speaking, given high-dimensional data $X = (x_1, \ldots, x_n)$, we want to estimate a low-dimensional model characterizing the population.
- Why do we care?
  - Simulation environments, prediction, inverse problems, transfer learning.
- **Problem**: X is itself high-dimensional! (unless you believe in the low-dimensionality manifold hypothesis)
- Prior can be encoded in a parametric generative model: density estimation.
  - Ex: GMM is a shallow model that assumes density concentrates in a finite number of modes
  - If data is sequential, exploit temporal regularity (eg word2vec).

Unsupervised Learning with Deep Networks

- How to learn a representation from unlabeled data that captures regularity AND complexity?

- How to relate auto-encoder models with variational inference?

Unsupervised Learning with Deep Networks

- How to learn a representation from unlabeled data that captures regularity AND complexity?
- How to relate auto-encoder models with variational inference?
- How to evaluate unsupervised models properly?
- How to relate deep representations with the method of moments and maximum entropy?

Optimization in Deep Learning

- Geoff Hinton, when describing their landmark 2012 Imagenet result: "We applied all the tricks that Yann and his lab had developed over the last 10 years…plus dropout"

Optimization in Deep Learning

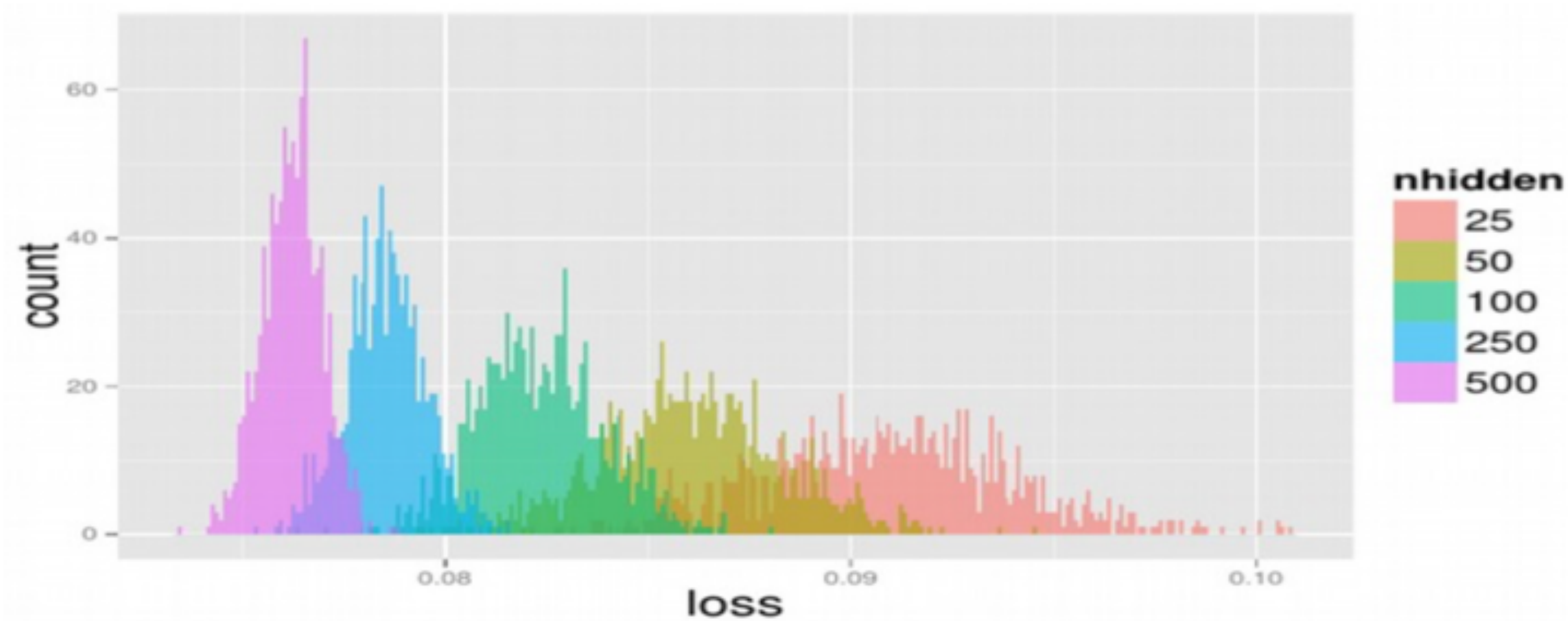
- Geoff Hinton, when describing their landmark 2012 Imagenet result: "We applied all the tricks that Yann and his lab had developed over the last 10 years…plus dropout"

- There is a functional equivalence between models of different depths at equal capacity ([Ba and Caruana'14]).

- So why deep models perform better in practice?

Non-Convex Optimization

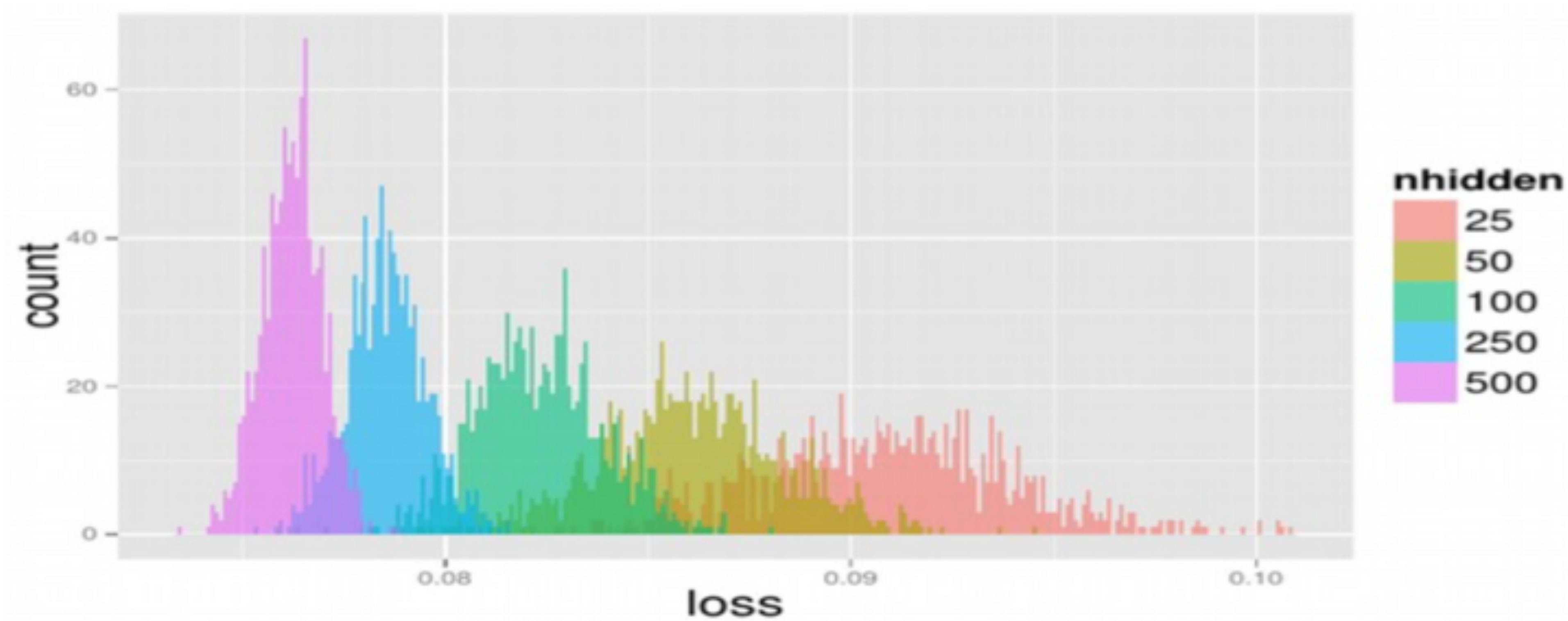
- [Choromaska et al, AISTATS'15] (also [Dauphin et al, ICML'15]) use tools from Statistical Physics to explain the behavior of stochastic gradient methods when training deep neural networks.



[Choromaska et al, AISTATS'15]

Non-Convex Optimization

- [Choromaska et al, AISTATS'15] (also [Dauphin et al, ICML'15]) use tools from Statistical Physics to explain the behavior of stochastic gradient methods when training deep neural networks.



[Choromaska et al, AISTATS'15]

- Offers a macroscopic explanation of why SGD "works".
- Gives a characterization of the network depth.
- Strong model simplifications, no convolutional specification.

Tentative Agenda

1. Convolutional and Recurrent Neural Networks
   – Invariance, Stability
   – Scattering Networks
   – Supervised Learning with CNNs
   – Properties of CNNs
   – Recurrent Models
   – *Guest Lecture:* **Wojciech Zaremba (OpenAI)**
2. Unsupervised Learning with Deep Networks
   – Auto encoders
   – Variational Autoencoders
   – Gibbs models
   – Generative Adversarial Networks
   – *Guest Lecture:* **Ian Goodfellow (Google Brain)**
3. Optimization & Misc
   – Dropout, Batch Normalization
   – Non-convex Optimization and Tensor Decompositions
   – Reasoning, Attention and Memory (time permitting)
   – *Guest Lecture:* **TBA**