POLITECNICO DI MILANO





Deep Learning – Introduction

Matteo Matteucci – matteo.matteucci@polimi.it

Disclaimer and credits

The following lecture has been taken from the following sources:

- Andrew Ng, *Learning feature hierarchies and deep learning*, ECCV'10
- Yan LeCun, Marc'Aurelio Ranzato, *Deep Learning Tutorial*, ICML'13
- Honglak Lee, *Tutorial on Deep Learning and Applications*, NIPS'10
- Li Deng, *Deep Learning from Speech Analysis/Recognition to Language/Multimodal Processing*, ICML'14.

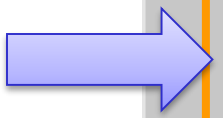
Big players in the field (beside Accademia)

- Microsoft
- Google
- Bing
- Facebook

Big names: Yoshua Bengio, Geoff Hinton, Yann LeCun, Marc'Aurelio Ranzato, Andrew Ng, …

MIT Technology Review

10 BREAKTHROUGH TECHNOLOGIES 2013



Deep Learning

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart.

Temporary Social Media

Messages that quickly self-destruct could enhance the privacy of online communications and make people freer to be spontaneous.

Prenatal DNA Sequencing

Reading the DNA of fetuses will be the next frontier of the genomic revolution. But do you really want to know about the genetic problems or musical aptitude of your unborn child?

Additive Manufacturing

Skeptical about 3-D printing? GE, the world's largest manufacturer, is on the verge of using the technology to make jet parts.

Baxter: The Blue-Collar Robot

Rodney Brooks's newest creation is easy to interact with, but the complex innovations behind the robot show just how hard it is to get along with people.

Memory Implants

A maverick neuroscientist believes he has deciphered the code by which the brain forms long-term memories. Next: testing a prosthetic implant for people suffering from long-term memory loss.

Smart Watches

The designers of the Pebble watch realized that a mobile phone is more useful if you don't have to take it out of your pocket.

Ultra-Efficient Solar Power

Doubling the efficiency of a solar cell would completely change the economics of renewable energy. Nanotechnology just might make it possible.

Big Data from Cheap Phones

Collecting and analyzing information from simple cell phones can provide surprising insights into how people move about and behave – and even help us understand the spread of diseases.

Supergrids

A new high-power circuit breaker could finally make highly efficient DC power grids practical.

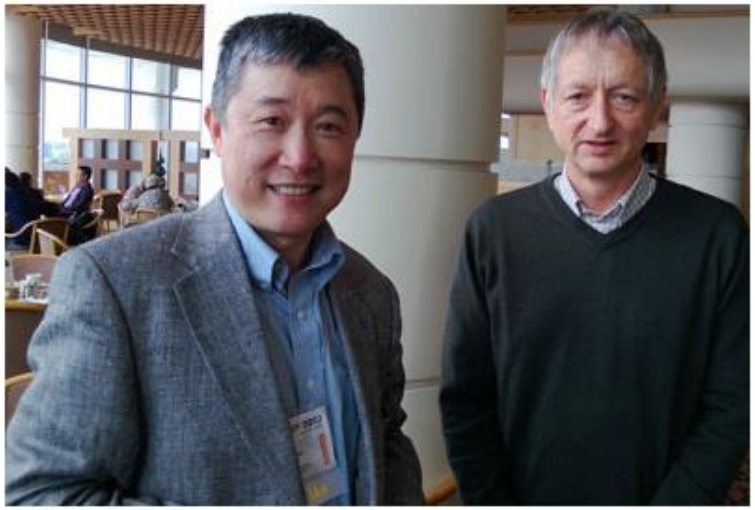
Enabling Cross-Lingual Conversations in Real Time

Microsoft Research
May 27, 2014 5:58 PM PT

The success of the team's progress to date was on display May 27, in a talk by Microsoft CEO Satya Nadella in Rancho Palos Verdes, Calif., during the Code Conference. During Nadella's conversation with Kara Swisher and Walt Mossberg of the Re/code tech website relating to a new

The path to the Skype Translator gained momentum with an encounter in the autumn of 2010. Seide and colleague Kit Thambiratnam had developed a system they called The Translating! Telephone for live speech-to-text and speech-to-speech trans calls.

e asked
tage. Pall, the
ident of
first time
app, with Pall
rman-
Diana



Li Deng (*left*) and Geoff Hinton.

A core development that enables Skype translation came from Redmond researcher Li Deng. He invited Geoff Hinton, a professor at the University of Toronto, to visit Redmond in 2009 to work on new neural-network learning methods, based on a couple of seminal papers from Hinton and his collaborators in 2006 that had brought new

/ MOBILE

Microsoft's Skype "Star Trek" Language Translator Takes on Tower of Babel

May 27, 2014, 5:48 PM PDT

Remember the universal translator on Star Trek? The gadget that let Kirk and Spock talk to aliens?

Skype to get 'real-time' translator

Analysts say the translation feature could have wide ranging applications

MIT Technology Review

Facebook Launches Advanced AI Effort to Find Meaning in Your Posts

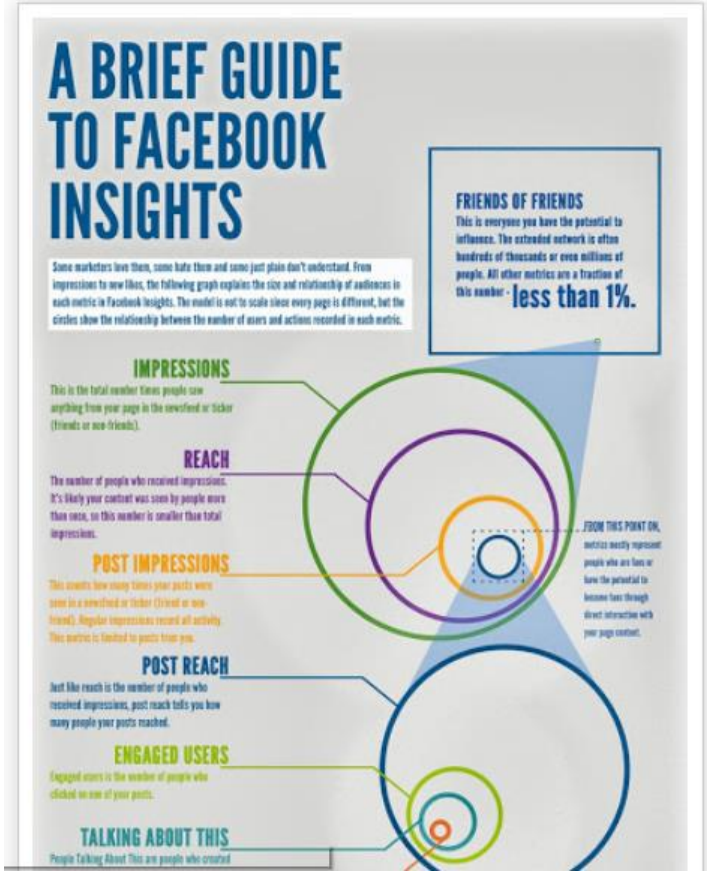
September 20, 2013

A technique called deep learning could help Facebook understand its users and their data better.

……Facebook's foray into deep learning sees it following its competitors Google and Microsoft, which have used the approach to impressive effect in the past year. Google has hired and acquired leading talent in the field (see "10 Breakthrough Technologies 2013: Deep Learning"), and last year created software that taught itself to recognize cats and other objects by reviewing stills from YouTube videos. The underlying deep learning technology was later used to slash the error rate of Google's voice recognition services (see "Google's Virtual Brain Goes to Work")

…. Researchers at Microsoft have used deep learning to build a system that translates speech from English to Mandarin Chinese in real time (see "Microsoft Brings Star Trek's Voice Translator to Life").

Chinese Web giant Baidu also recently established a Silicon Valley research lab to work on deep learning.



BloombergBusinessweek
Technology

Acquisitions

The Race to Buy the Human Brains Behind Deep Learning Machines

By Ashlee Vance | January 27, 2014

intelligence projects. “DeepMind is bona fide in terms of its research capabilities and depth,” says Peter Lee, who heads Microsoft Research.

According to Lee, Microsoft, Facebook (FB), and Google find themselves in a battle for deep learning talent. Microsoft has gone from four full-time deep learning experts to 70 in the past three years. “We would have more if the talent was there to be had,” he says. “Last year, the cost of a top, world-class deep learning expert was about the same as a top NFL quarterback prospect. The cost of that talent is pretty remarkable.”

Chinese Search Giant Baidu Hires Man Behind the "Google Brain"

China's Baidu Bets on Deep Learning

MIT Technology Review - 3 days ago

Deep learning makes it possible for machines to process large amounts of date using simulated networks of simple neurons, crudely modeled ...

Baidu snatches Google's deep-learning visionary, Andrew ...

DARPA is working on its own deep-learning project for natural-language processing

by Derrick Harris MAY. 2, 2014 - 10:49 AM PDT

SUMMARY: *The Defense Advanced Research Projects Agency, or DARPA, is building a set of technologies to help it better understand human language so it can analyze speech and text sources and alert analysts of potentially useful information*

artificial intelligence / machine-learning / open source

A startup called Skymind launches, pushing open source deep learning

by Derrick Harris JUN. 2, 2014 - 10:03 AM PDT

SUMMARY: *Skymind is providing commercial support and services for an open source project called deeplearning4j. It's a collection of of approaches to deep learning that mimic those developed by leading researchers, but tuned for enterprise adoption.*

photo: deviantART / rajasegar

POLITECNICO DI MILANO





Deep Learning – Herarchical feature learning

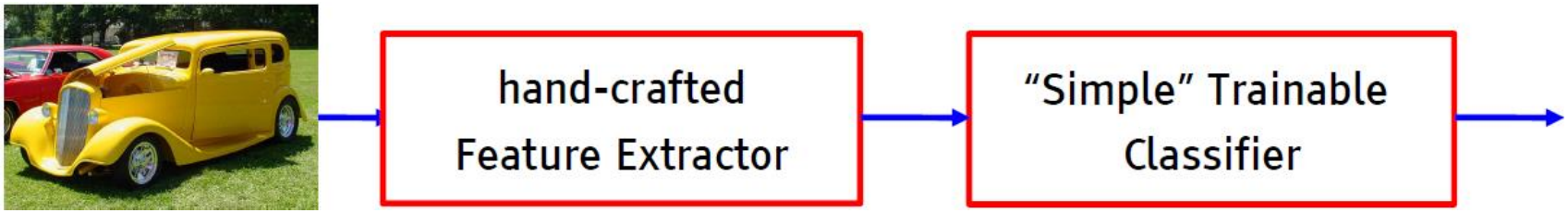
Matteo Matteucci – matteo.matteucci@polimi.it

Deep Learning in a nutshell

Yan LeCun, Marc'Aurelio Ranzato
(IMCL 2013)

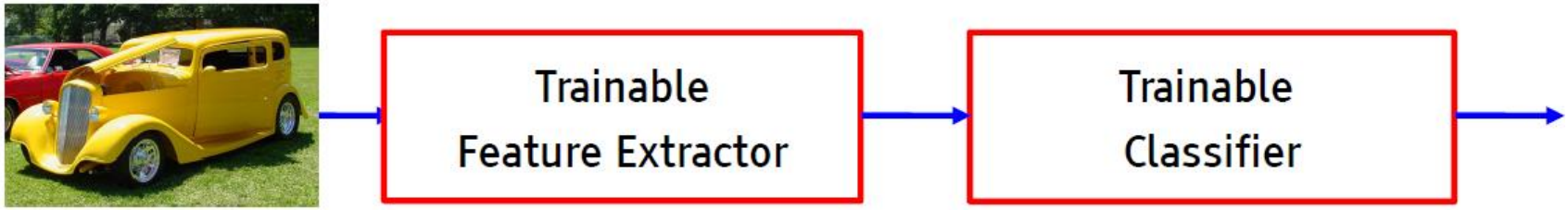
Deep learning can be summarized as learning both the representation and the classifier out of it

- Fixed engineered features (or kernels) + trainable classifier



VS.

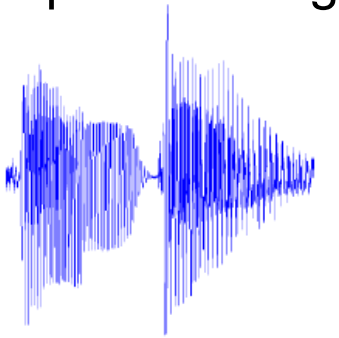
- End-to-end learning / feature learning / deep learning



Modern pattern recogniton

Yan LeCun, Marc'Aurelio Ranzato
(IMCL 2013)

Speech recognition (early 90's – 2011)



MFCC → Mix of Gaussians → Classifier

fixed — unsupervised — supervised

Low level features — Mid level features

Object recognition (2006 – 2012)

SIFT HoG → K-means Sparse Coding → Pooling → Classifier

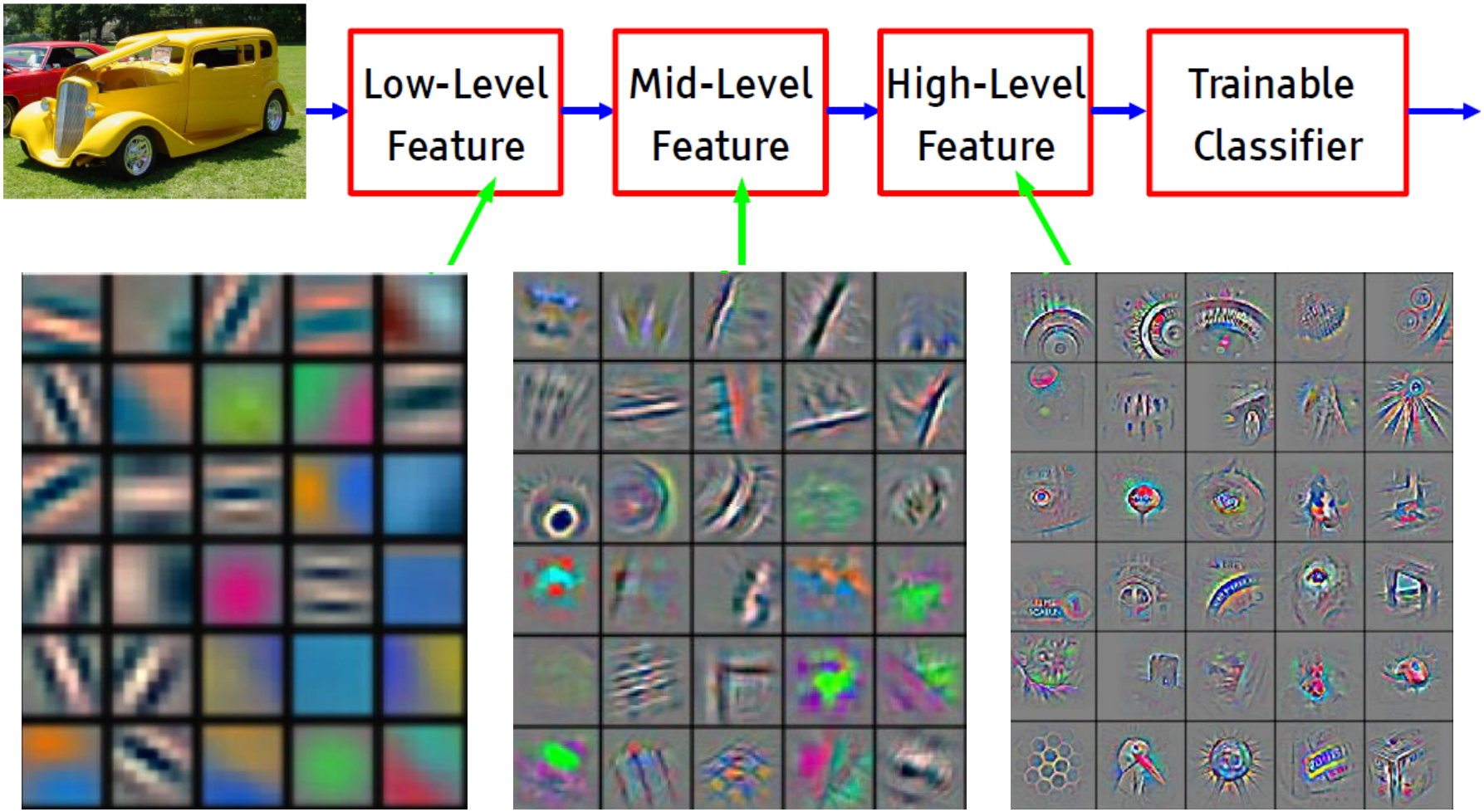
fixed — unsupervised — supervised

Low level features — Mid level features

Deep learning pipeline

Yan LeCun, Marc'Aurelio Ranzato
(IMCL 2013)

In deep learning we have multiple stages of non linear feature transformation



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Learning the representation

Yan LeCun, Marc'Aurelio Ranzato (IMCL 2013)

Learning the representation is a challenging problem for ML, CV, AI, Neuroscience, Cognitive Science, ...

Cognitive perspective

- How can a perceptual system build itself by looking at the external world?
- How much prior structure is necessary?

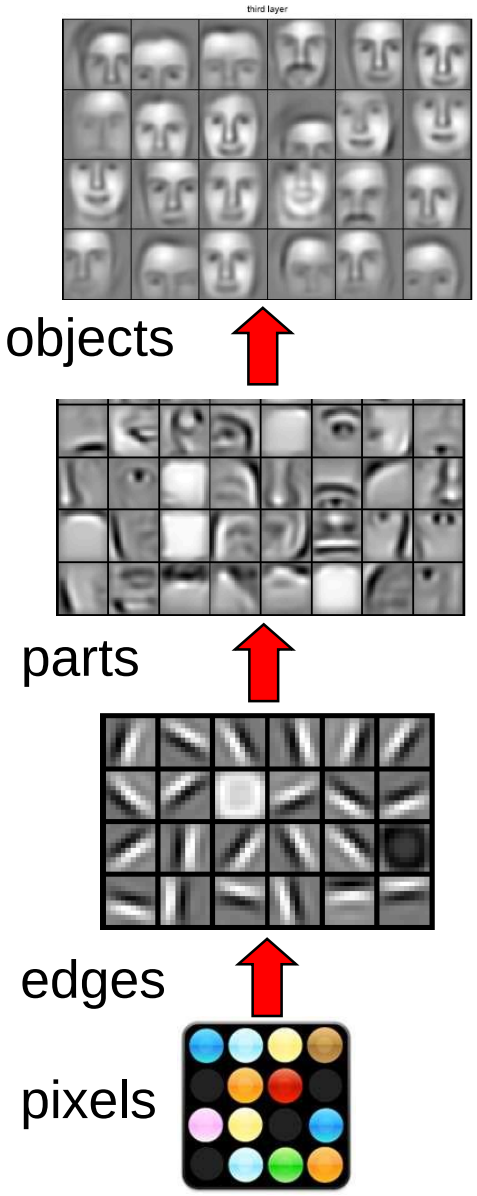
Neuroscience

- Does the cortex «run» a single, general learning algorithm? Or multiple simpler ones?

ML/AI Perspective

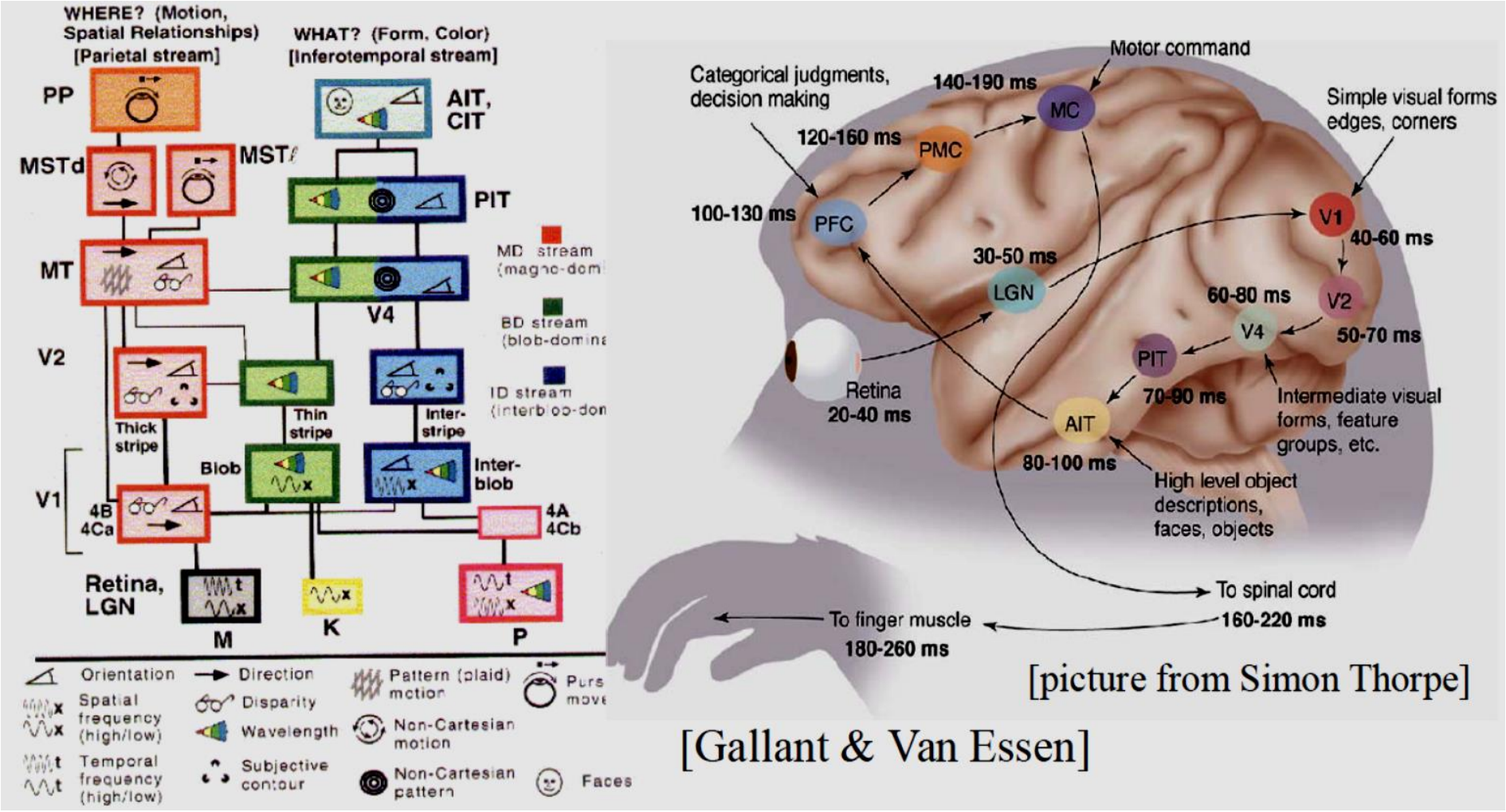
- What is the fundamental principle?
- What is the learning algorithm?
- What is the architecture?

Deep learning addresses the problem of learning hierarchical representations with a single algorithm



Mammalian visual cortex

The ventral (recognition) pathway in the visual cortex has multiple stages



Trainable features hierarchy

Yan LeCun, Marc'Aurelio Ranzato (IMCL 2013)

Deep learning assumes it is possible to «learn» a hierarchy of descriptors with increasing abstraction, i.e., layers are trainable feature transforms

In image recognition

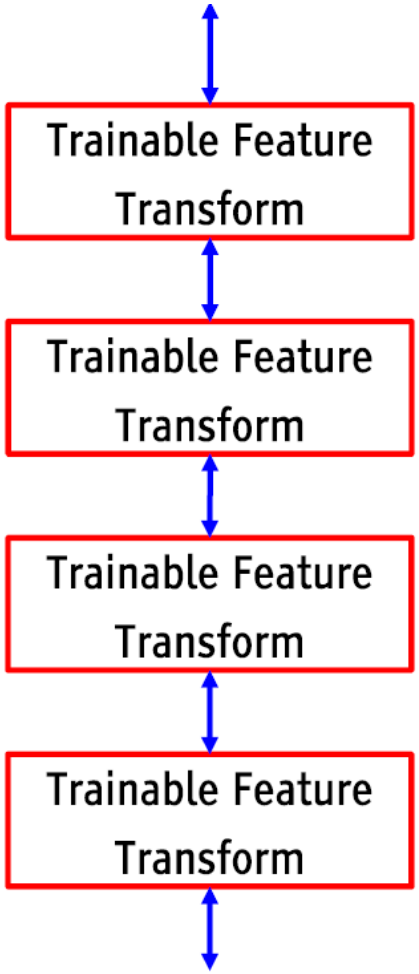
- Pixel → edge → texton → motif → part → → object

In text analysis

- Character → word → word group → clause → → sentence → story

In speech recognition

- Sample → spectral band → sound → ... → phone → phoneme → word

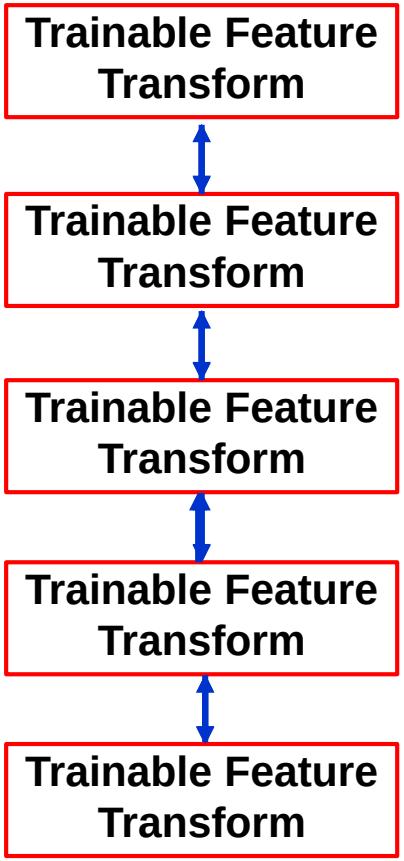


Architectures and algorithms

Yan LeCun, Marc'Aurelio Ranzato (IMCL 2013)

Depending on the direction of the information flow we can have different architectures for the hierarchy of features

- Feed forward (e.g., multilayer neural nets, convolutional nets, etc.)
- Feed back (e.g., Stacked sparse coding, Deconvolutional nets, etc.)
- Bi-directional (e.g, Deep Boltmann Machines, Stacked Auto-Encoders, etc.)

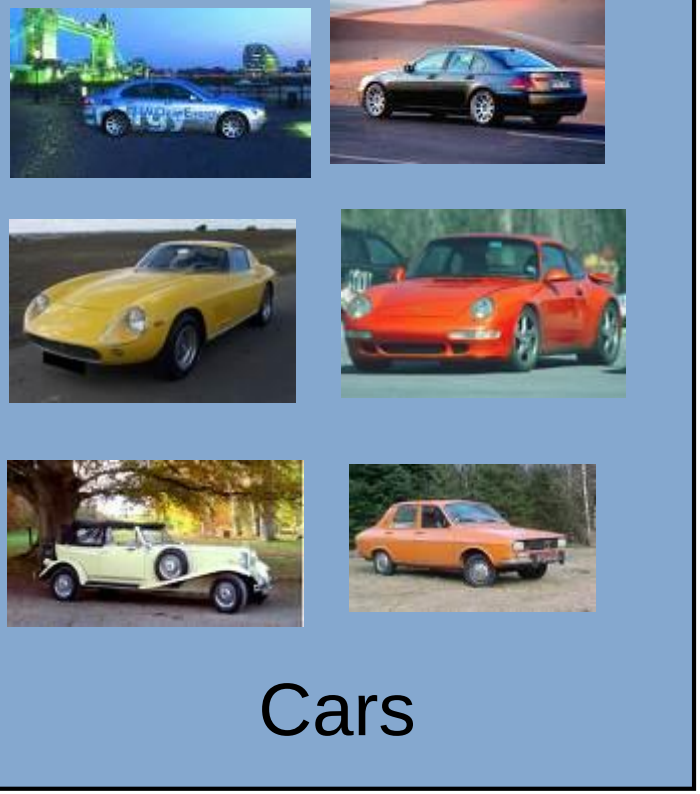


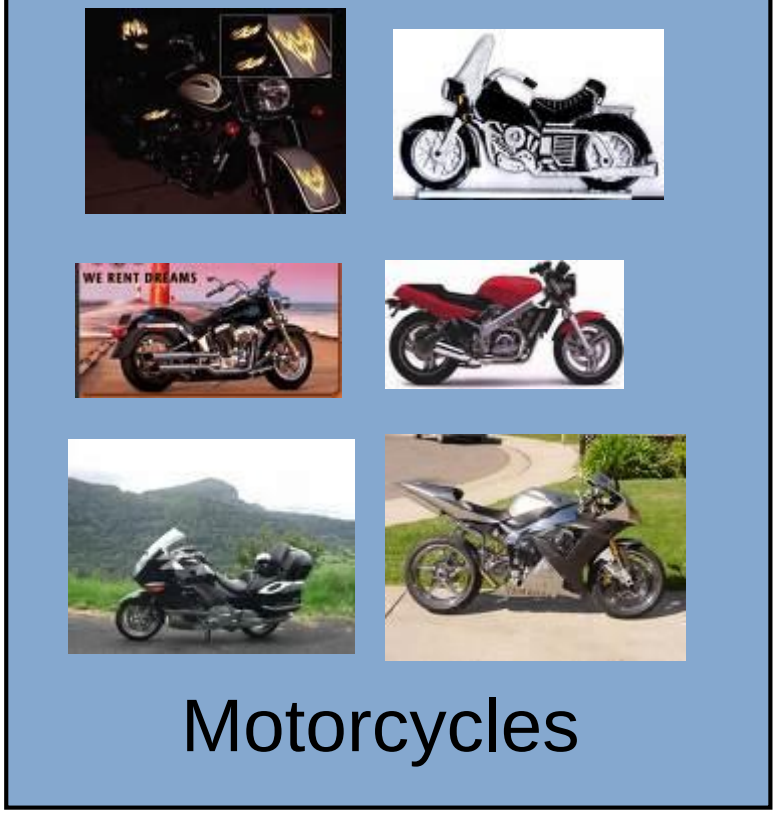
We can have also different kind of learning protocols

- Purely supervised
- Unsupervised (layerwise) + supervised on top
- Unsupervised (layerwise) + global supervised
- Unsupervised pre-training through regularized auto-encoders + ...

Supervised learning

Kay Yu, Andrew Ng
(ECCV 2010)



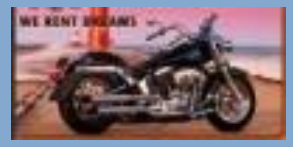


Testing:
What is this?

Semi-supervised learning

Kay Yu, Andrew Ng
(ECCV 2010)







Unlabeled images (all cars/motorcycles)

Car

Motorcycle

Testing:
What is this?

Self-taught learning

Kay Yu, Andrew Ng
(ECCV 2010)

Unlabeled images (random internet images)

Car

Motorcycle

Testing:
What is this?

Self-taught learning

Kay Yu, Andrew Ng
(ECCV 2010)

Unlabeled images
(random internet images)

→ Sparse coding, LCC, etc. → $\phi_1, \phi_2, \ldots, \phi_k$

Use learned $\phi_1, \phi_2, \ldots, \phi_k$ to represent training/test sets.

Car

Motorcycle

Using $\phi_1, \phi_2, \ldots, \phi_k$ → $a_1, a_2, \ldots, a_k$

When the labeled dataset is small can give significant improvement!

POLITECNICO DI MILANO





Deep Learning – Neural Networks Autoencoders

Matteo Matteucci – matteo.matteucci@polimi.it

Logistic regression

Kay Yu, Andrew Ng
(ECCV 2010)

Logistic regression has a learned parameter vector θ.

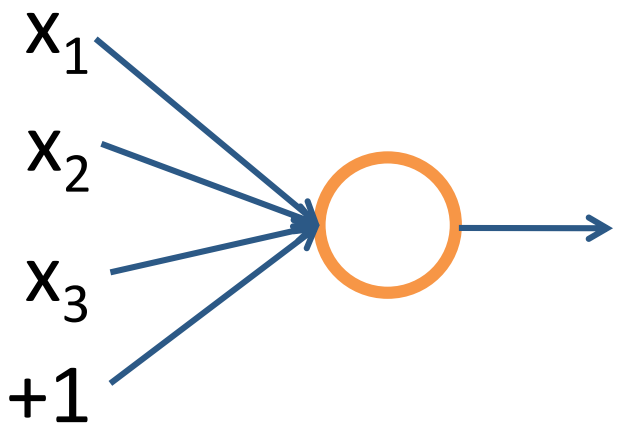
On input x, it outputs:

$$h_\theta(x) = \sigma(\theta^T x) = \frac{1}{1+\exp(-\theta^T x)}$$

where

$$\sigma(z) = 1/(1+\exp(-z))$$

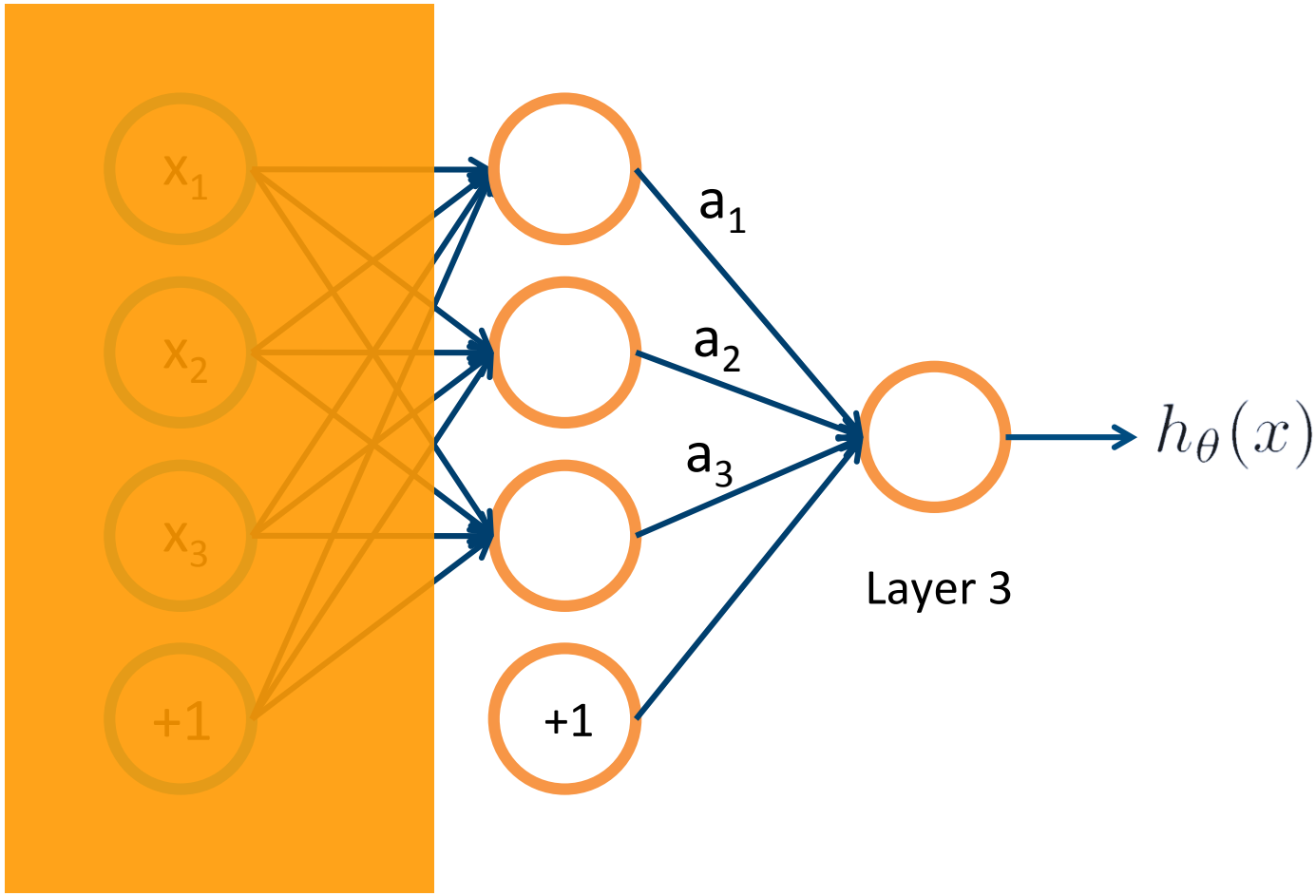
A logistic regression unit can be drawn as



$$h_\theta(x) = \frac{1}{1+\exp(-\theta^T x)}$$

Kay Yu, Andrew Ng
(ECCV 2010)

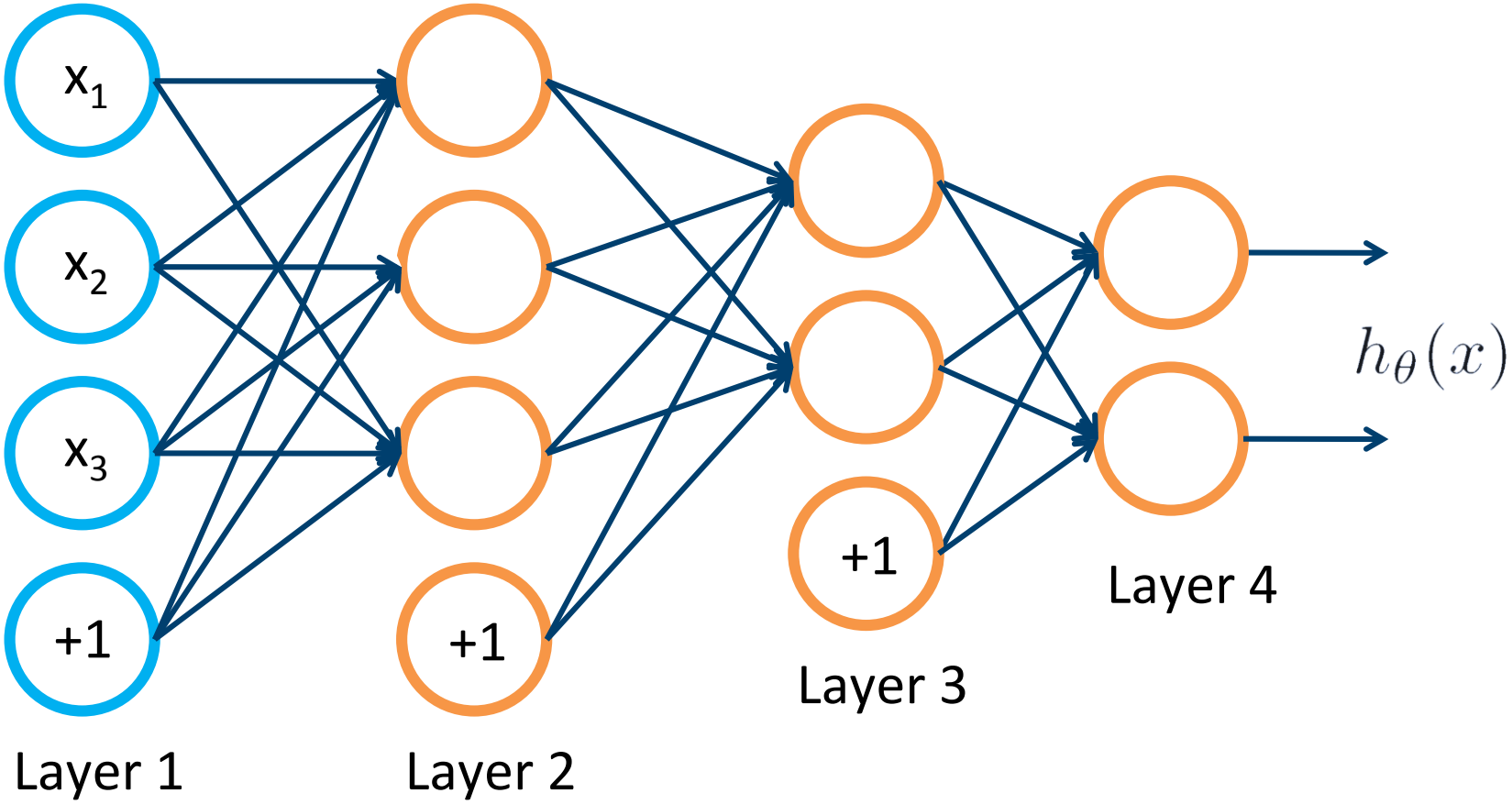
String a lot of logistic units together into a net (e.g., 3 layers)



Neural Network

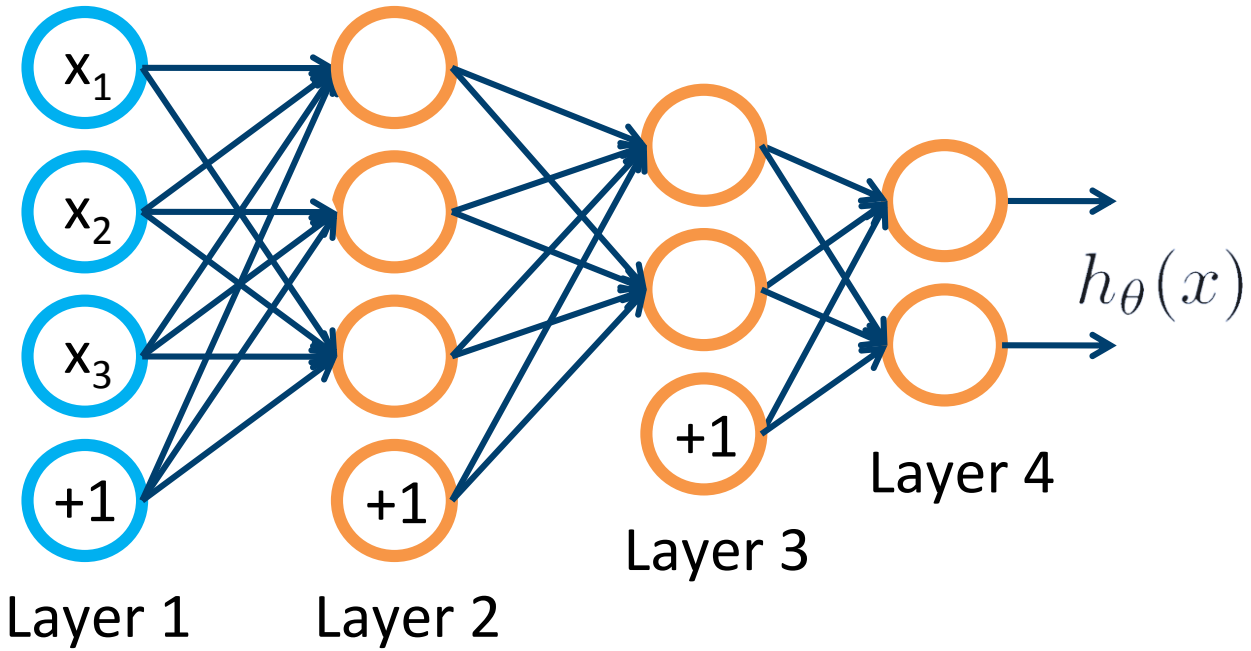
Kay Yu, Andrew Ng
(ECCV 2010)

Example 4 layer network with 2 output units:



Training a neural network

Kay Yu, Andrew Ng
(ECCV 2010)



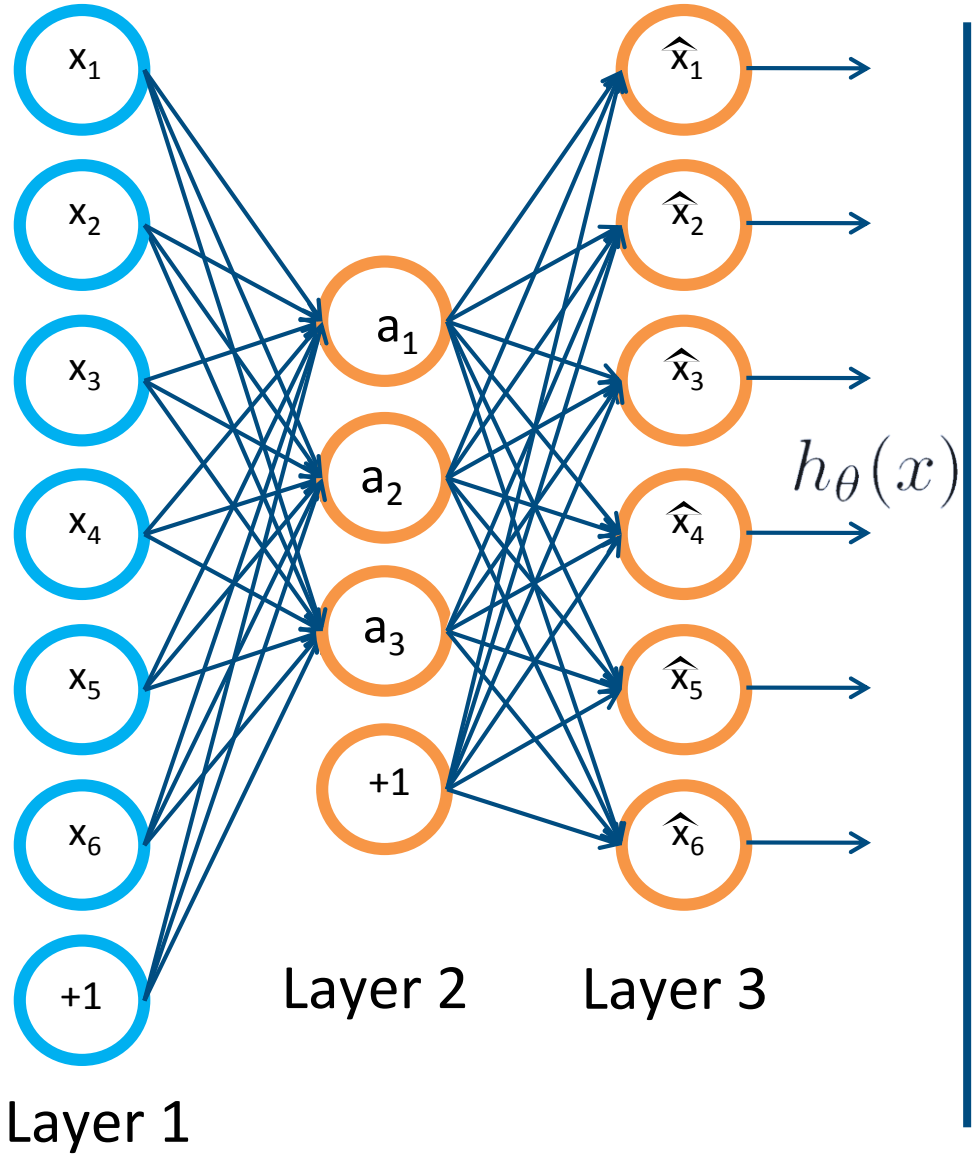
Given training set (x_1, y_1), (x_2, y_2), (x_3, y_3), …. adjust parameters θ (for every node) to make:

$$h_\theta(x_i) \approx y_i$$

using classic gradient descent (i.e., backpropagation) algorithm

Unsupervised feature learning

Kay Yu, Andrew Ng
(ECCV 2010)



Autoencoder.

Network trained to output the input (i.e., to learn the identity function).

$$h_\theta(x) \approx x$$

Trivial solution unless:

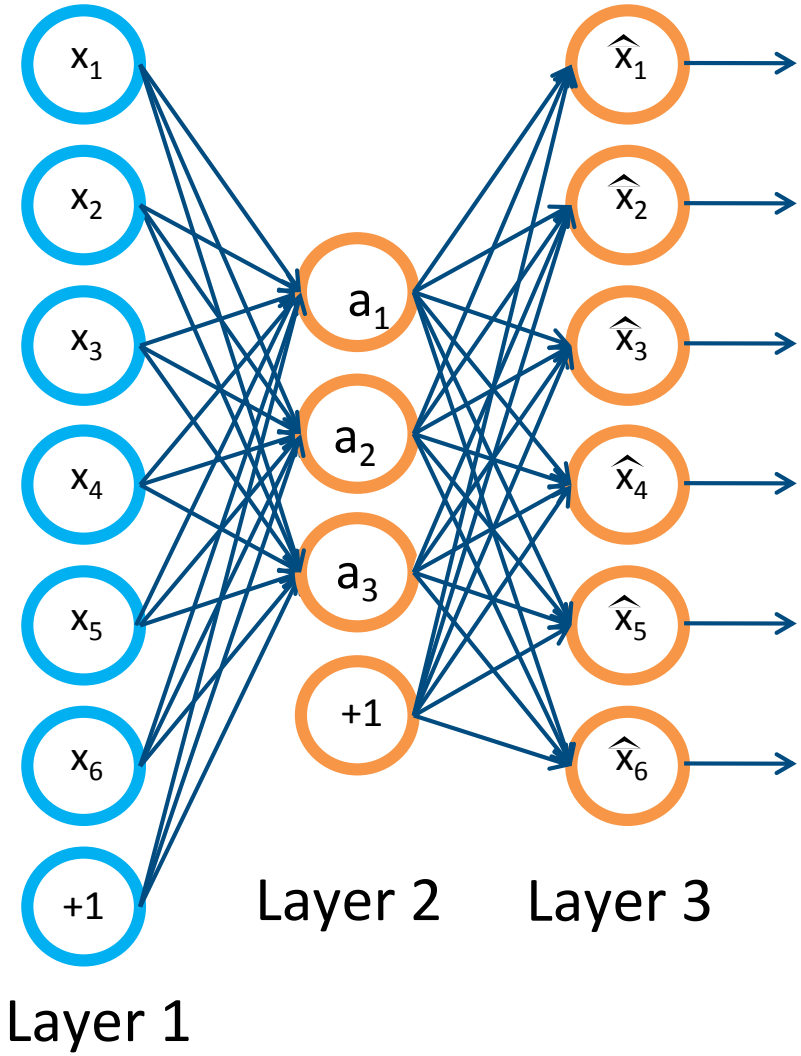
- Limited number of units in Layer 2 (compressed representation)
- Constrain Layer 2 to be **sparse** (sparse representation)

Unsupervised feature learning

Kay Yu, Andrew Ng
(ECCV 2010)

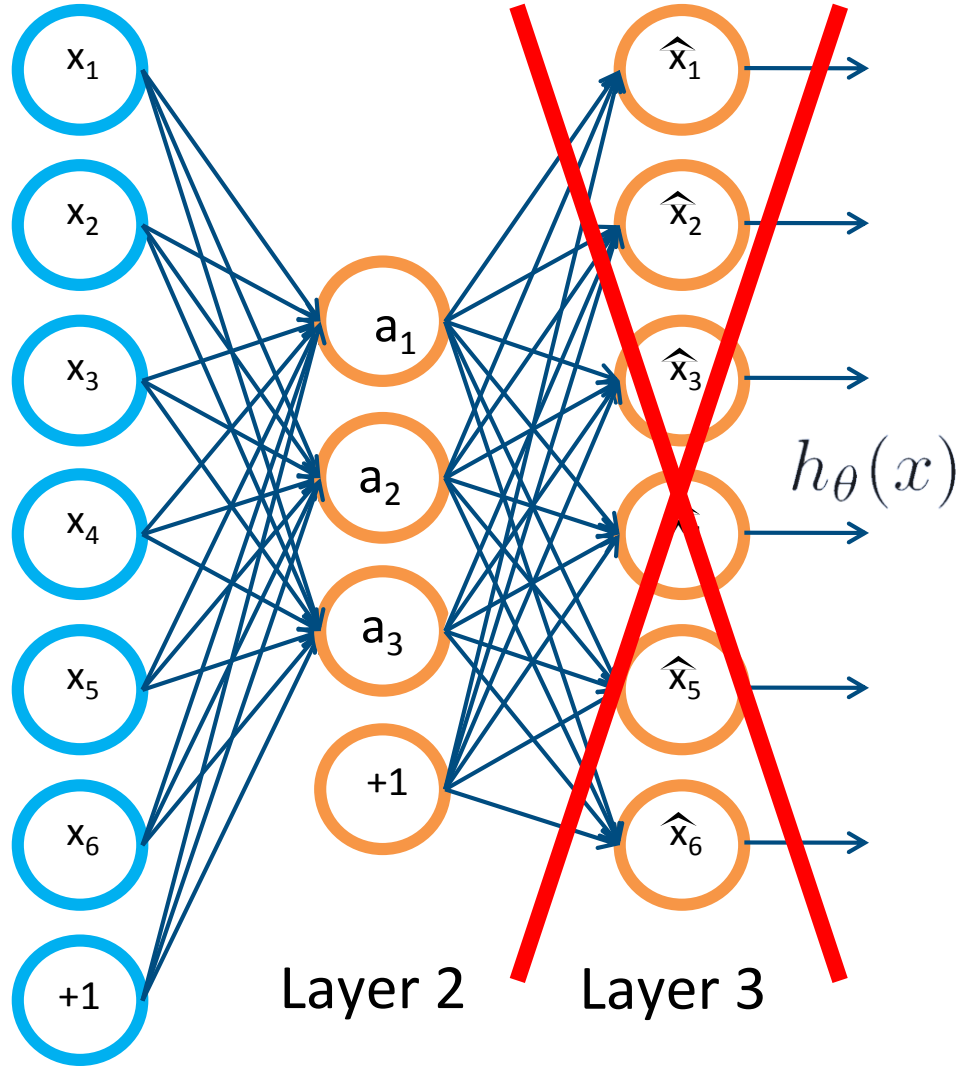
Training a sparse autoencoder from a dateset of unlabeled training samples x_1, x_2, … by

$$\min_{\theta} \underbrace{\|h_\theta(x) - x\|^2}_{\text{Reconstruction error term}} + \underbrace{\lambda \sum_i |a_i|}_{L_1 \text{ sparsity term}}$$



Neural Network Autoencoder

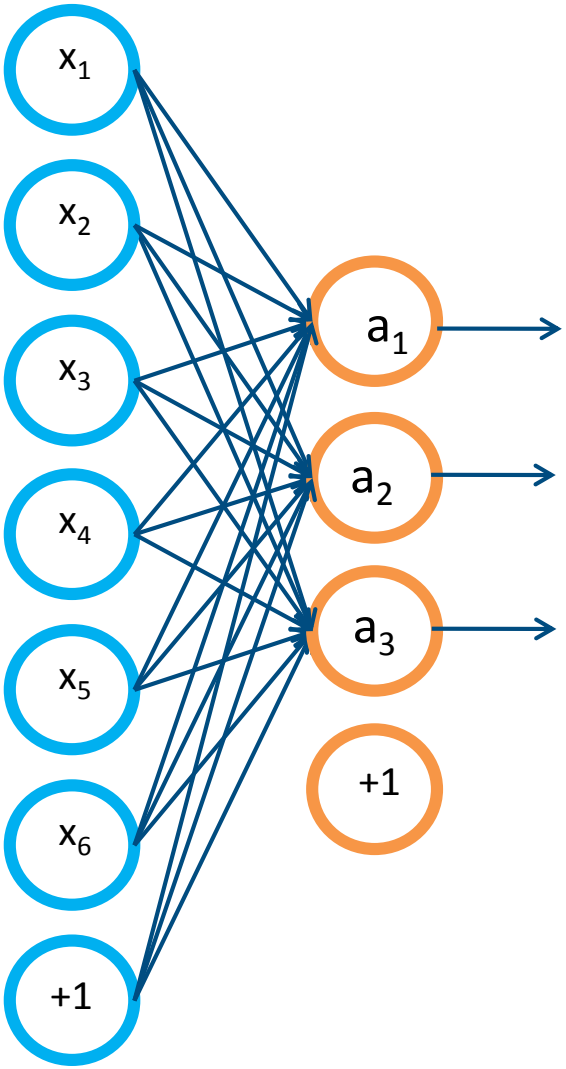
Kay Yu, Andrew Ng
(ECCV 2010)



Matteo Matteucci – matteo.matteucci@polimi.it

Neural Network Autoencoder

Kay Yu, Andrew Ng
(ECCV 2010)



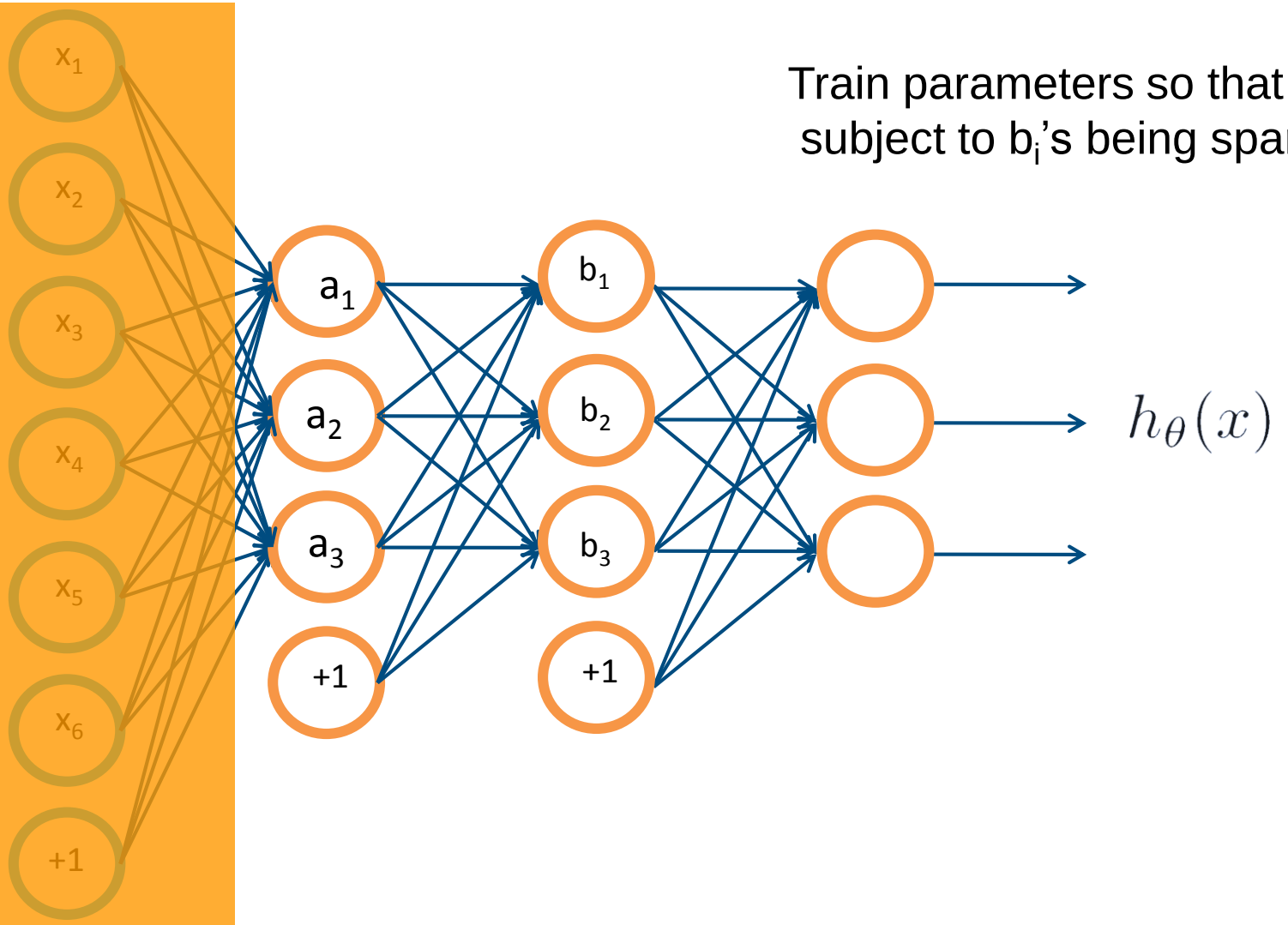
$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

New representation for the input.

Neural Network Autoencoder

Kay Yu, Andrew Ng
(ECCV 2010)

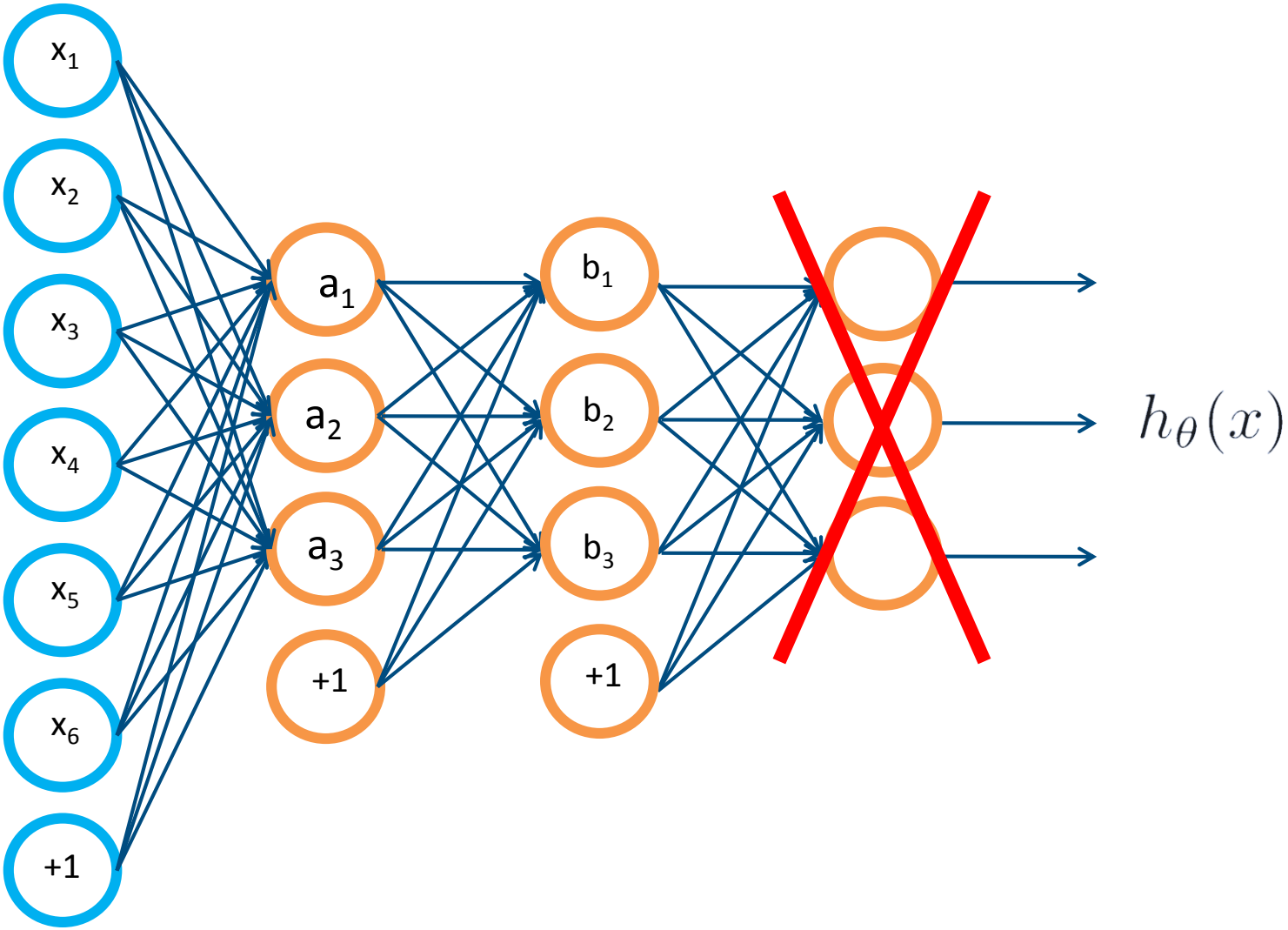
Train parameters so that $h_\theta(x) \approx a$, subject to b_i's being sparse.



Matteo Matteucci – matteo.matteucci@polimi.it

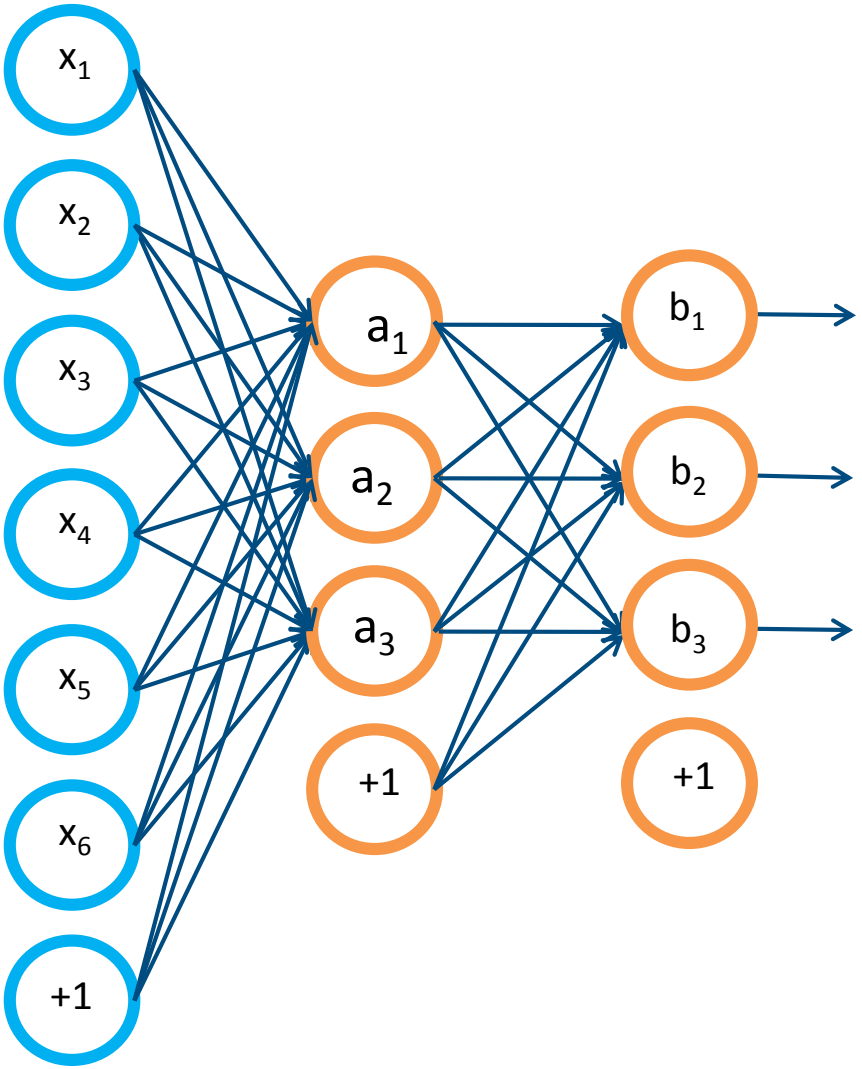
Neural Network Autoencoder

Kay Yu, Andrew Ng
(ECCV 2010)



Neural Network Autoencoder

Kay Yu, Andrew Ng
(ECCV 2010)



$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

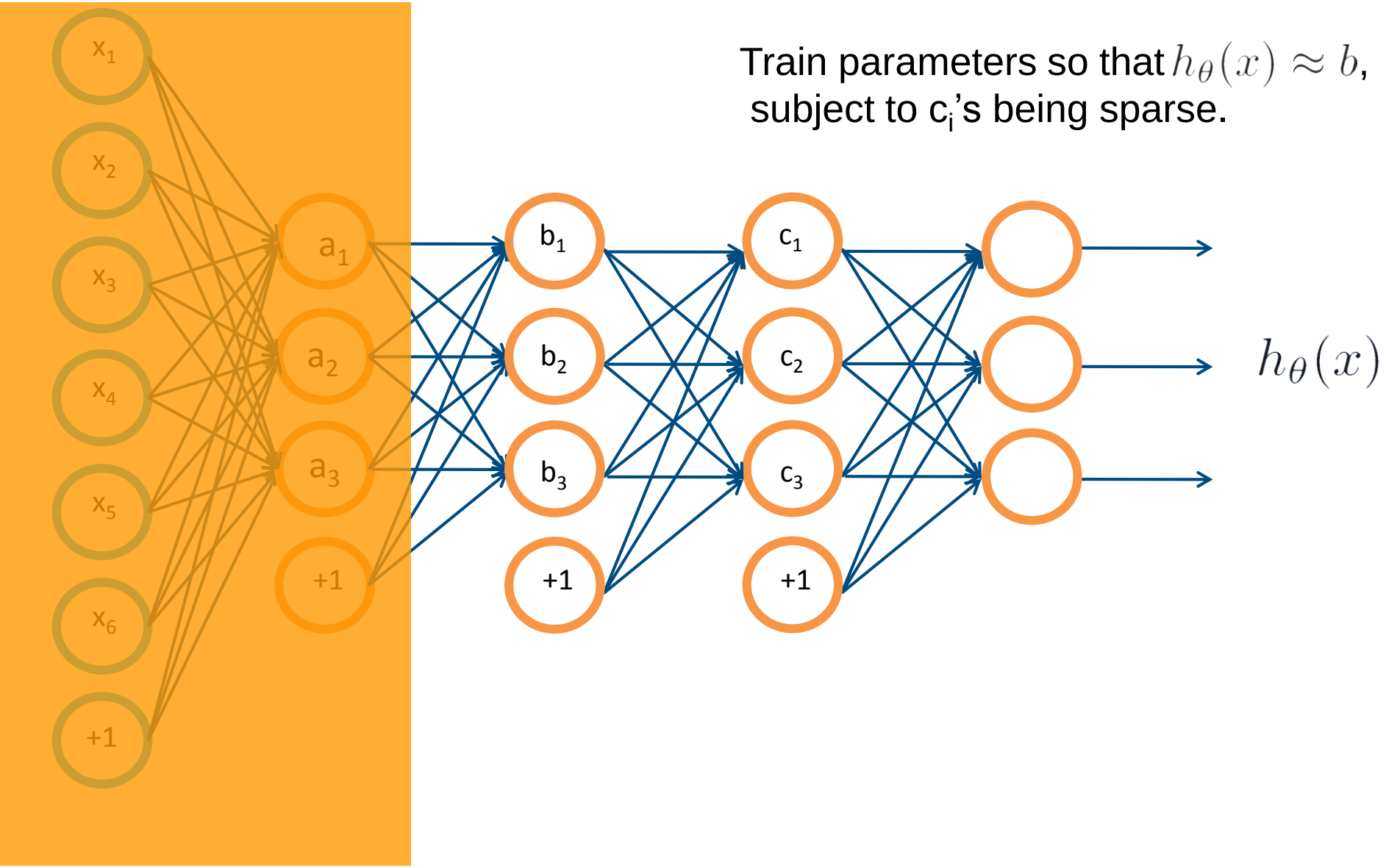
New representation for the input

Matteo Matteucci – matteo.matteucci@polimi.it

Neural Network Autoencoder

Kay Yu, Andrew Ng
(ECCV 2010)

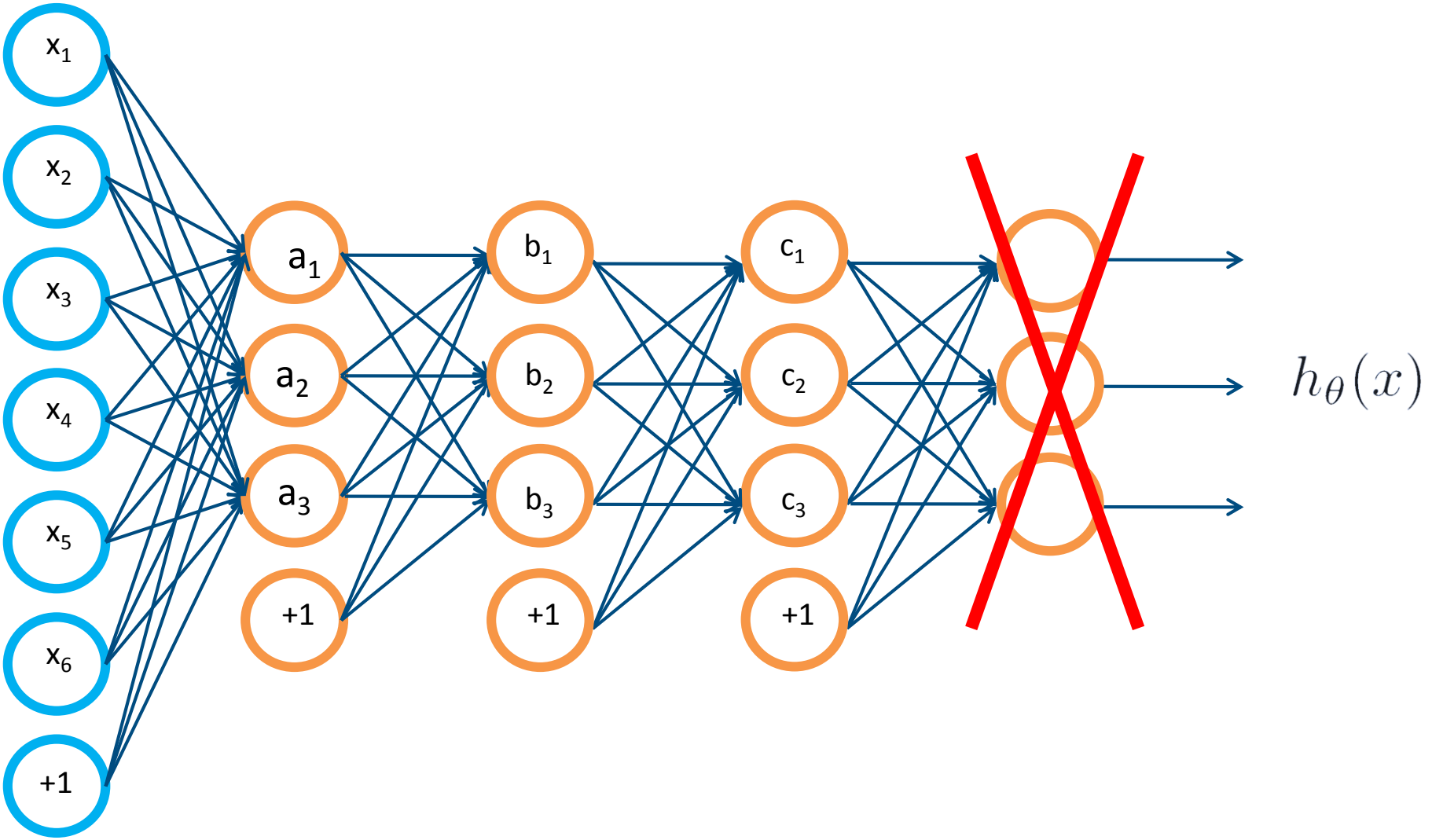
Train parameters so that $h_\theta(x) \approx b$, subject to c_i's being sparse.



Matteo Matteucci – matteo.matteucci@polimi.it

Neural Network Autoencoder

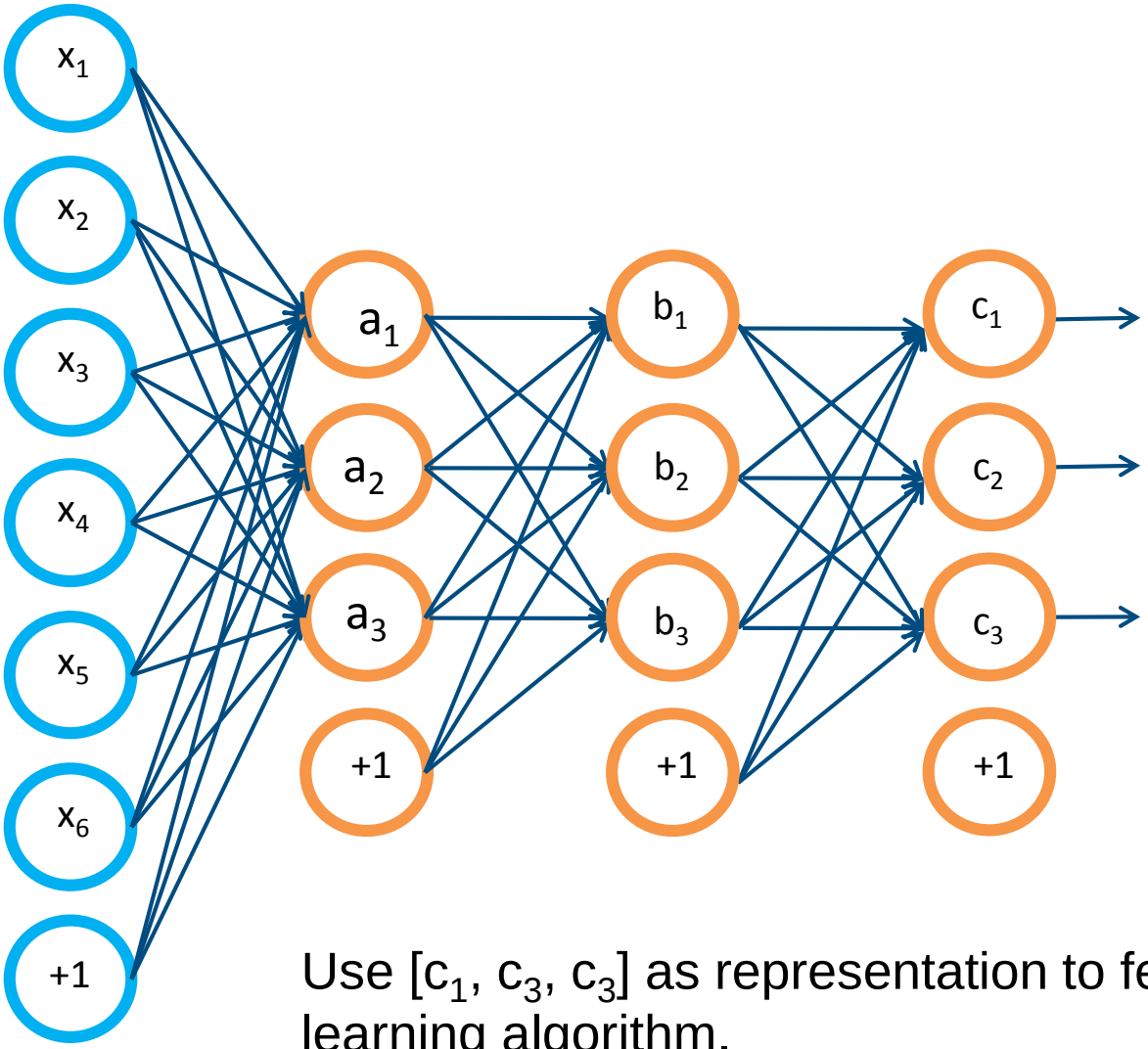
Kay Yu, Andrew Ng
(ECCV 2010)



Matteo Matteucci – matteo.matteucci@polimi.it

Neural Network Autoencoder

Kay Yu, Andrew Ng
(ECCV 2010)



$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$$

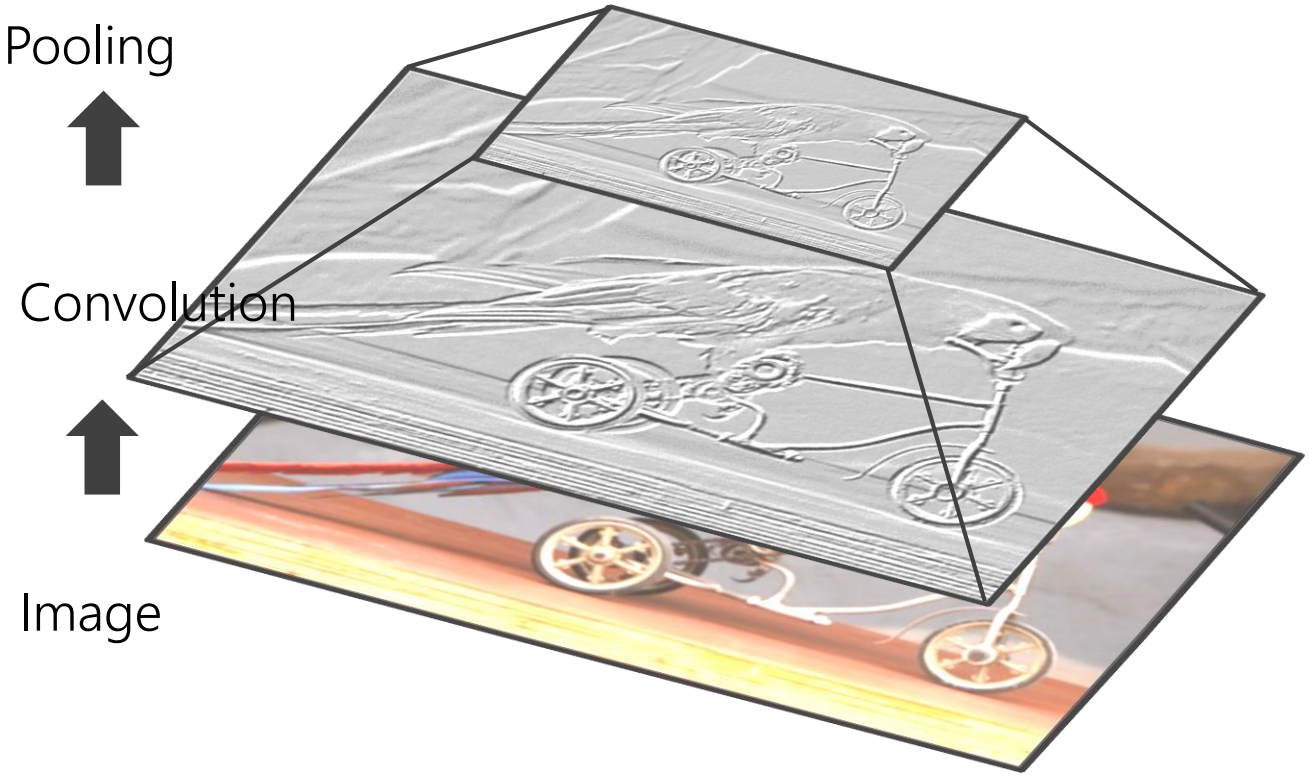
New representation for input.

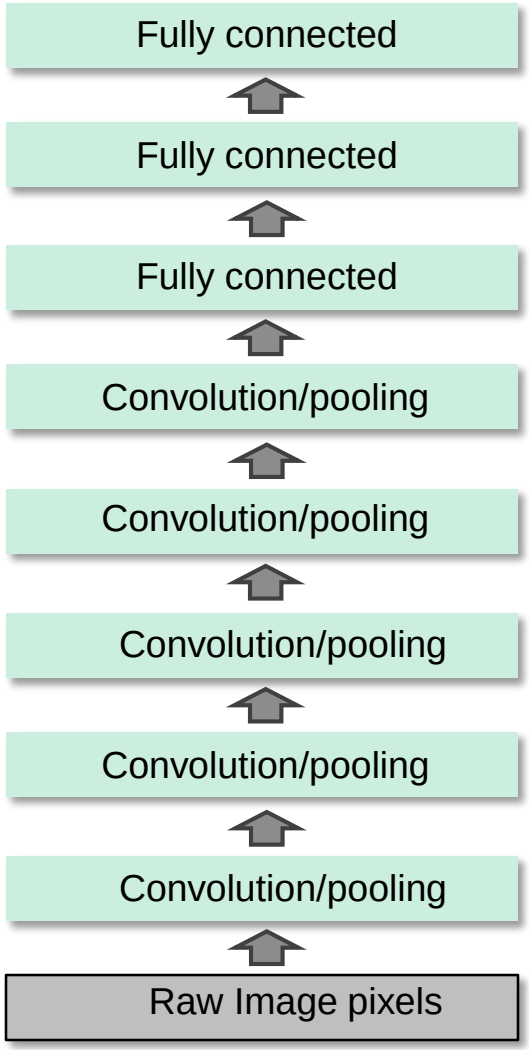
Use $[c_1, c_3, c_3]$ as representation to feed to a supervised learning algorithm.

Deep Convolutional NN

Li Deng
(IMCL 2014)

CNN: local connections with weight sharing;
pooling for translation invariance

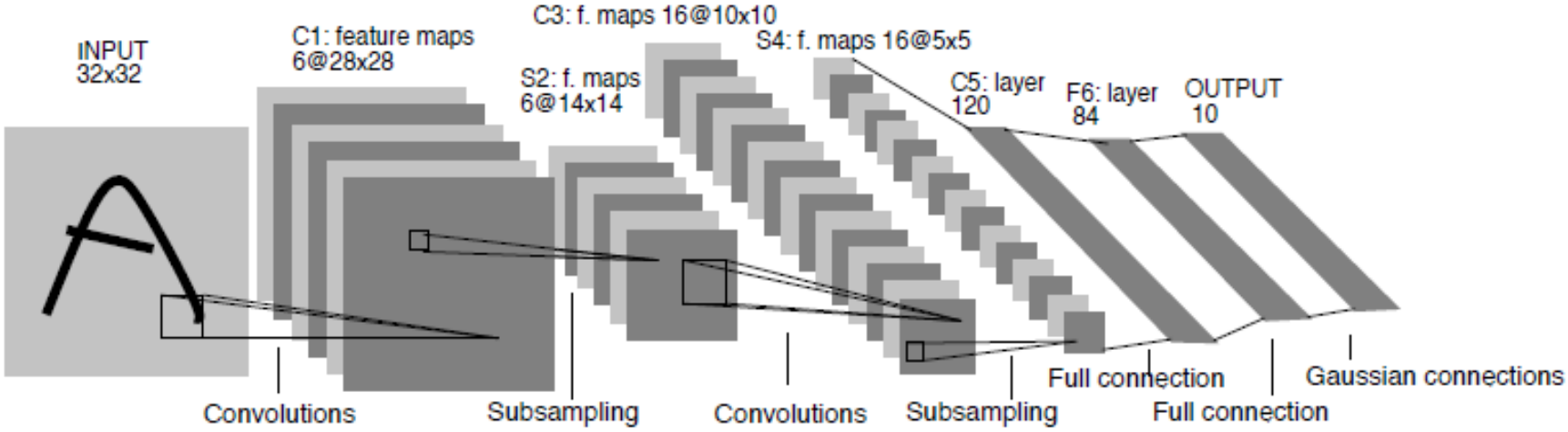




Deep CNN for image recognition

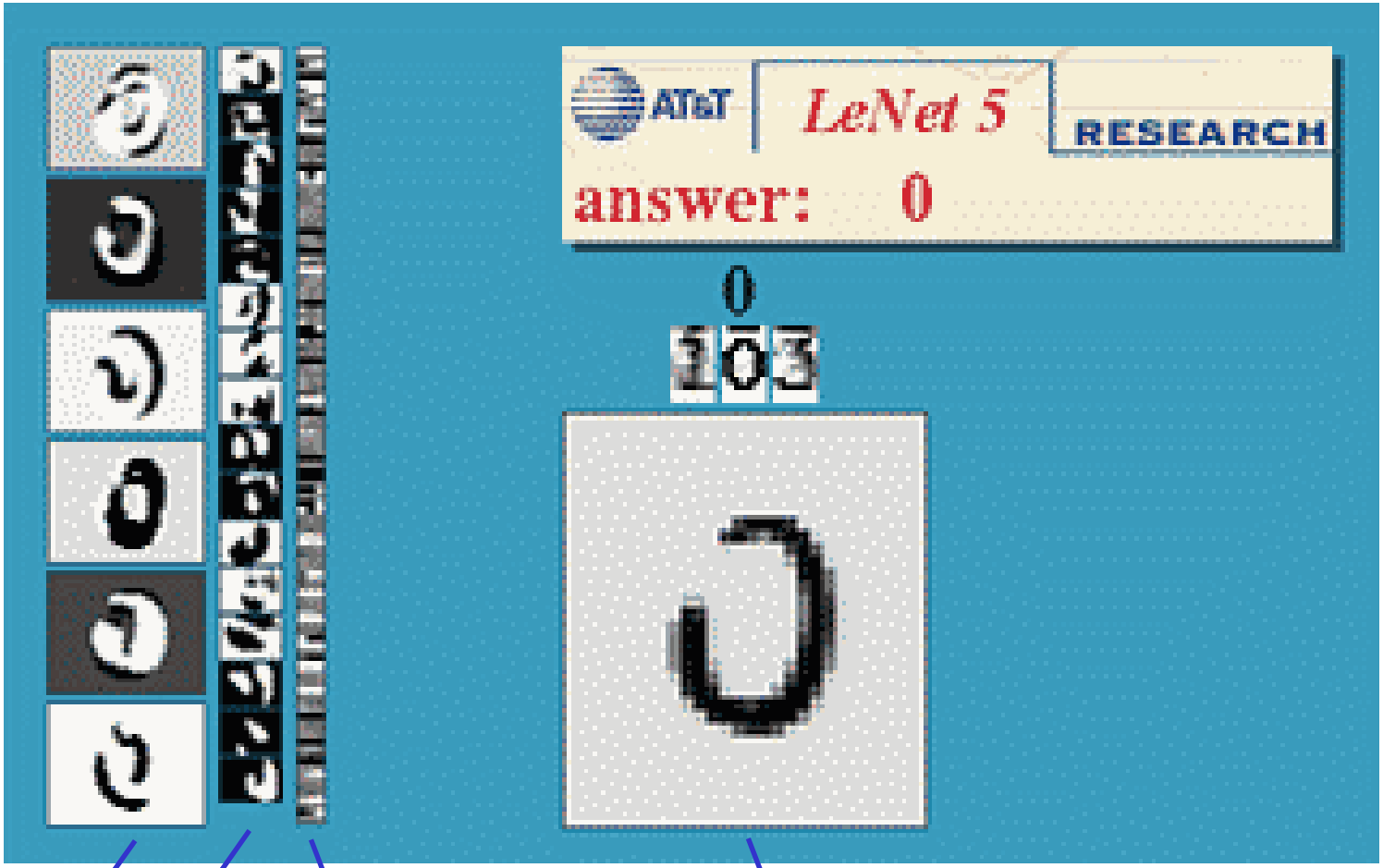
Li Deng
(IMCL 2014)

LeCun et al. 1998



LeNet-5

Kay Yu, Andrew Ng
(ECCV 2010)

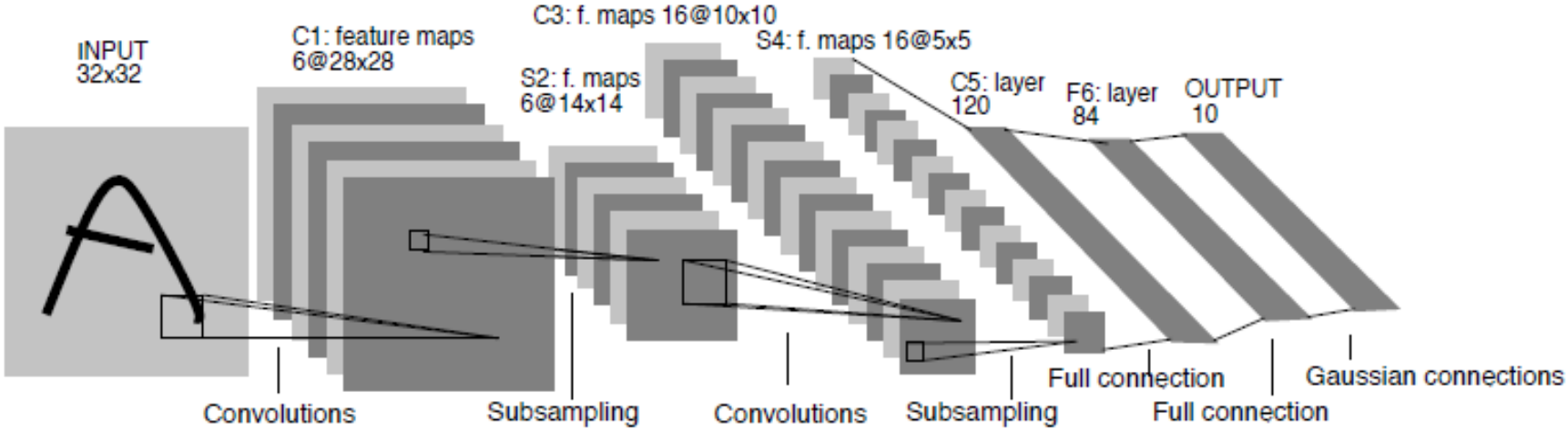


[Courtesy of Yan LeCun]

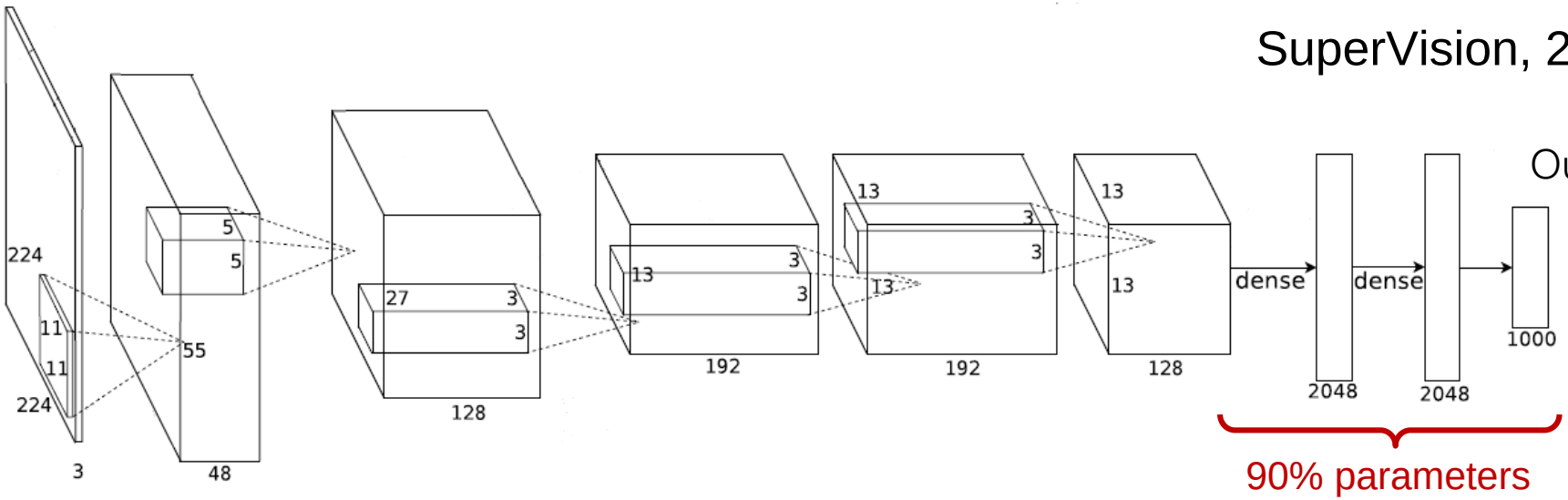
Deep CNN for image recognition

Li Deng
(IMCL 2014)

LeCun et al. 1998

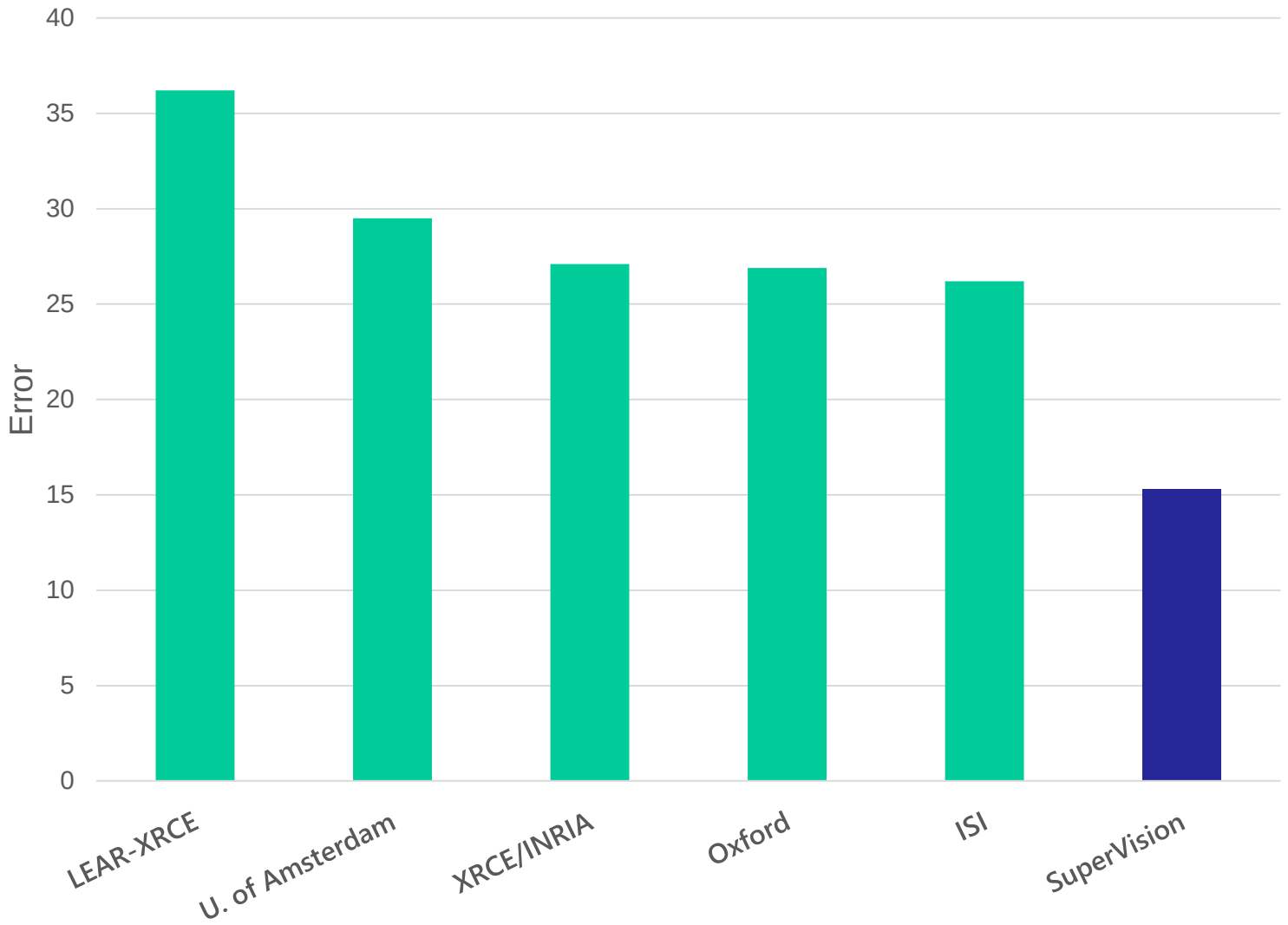


SuperVision, 2012



ImageNet 1K Competition (2012)

Li Deng
(IMCL 2014)



POLITECNICO DI MILANO





Deep Learning – Deep Belief Networks

Matteo Matteucci – matteo.matteucci@polimi.it

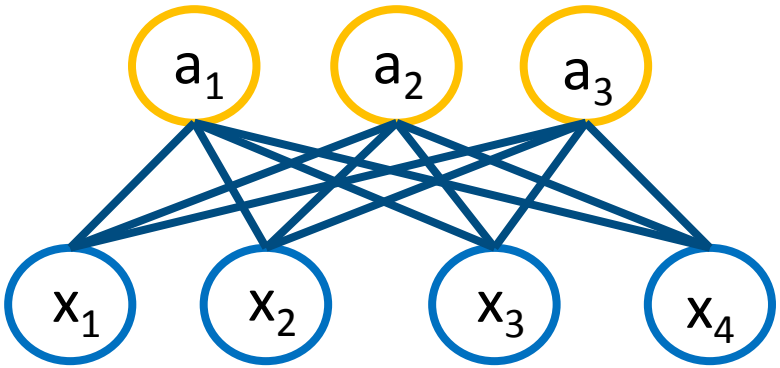
Deep Belief Net

Kay Yu, Andrew Ng
(ECCV 2010)

Deep Belief Net (DBN) is another algorithm for learning a feature hierarchy based on probabilistic graphical models.

Building block:

- 2-layer graphical model (Restricted Boltzmann Machine).



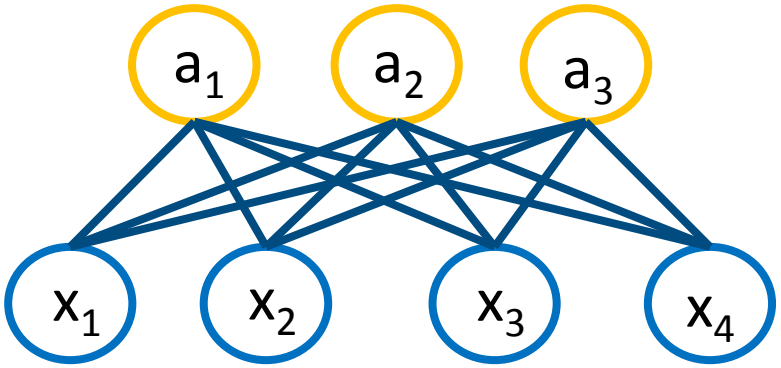
Layer 2. $[a_1, a_2, a_3]$ (binary-valued)

Input $[x_1, x_2, x_3, x_4]$

- We can learn additional layers one at a time
- We can add a supervised classifier at the end

Restricted Boltzmann machine (RBM)

Kay Yu, Andrew Ng
(ECCV 2010)



Layer 2. [a_1, a_2, a_3] (binary-valued)

Input [x_1, x_2, x_3, x_4]

MRF with joint distribution:

$$P(x,a) \propto \exp\left(-\sum_{i,j} x_i a_j W_{ij}\right)$$

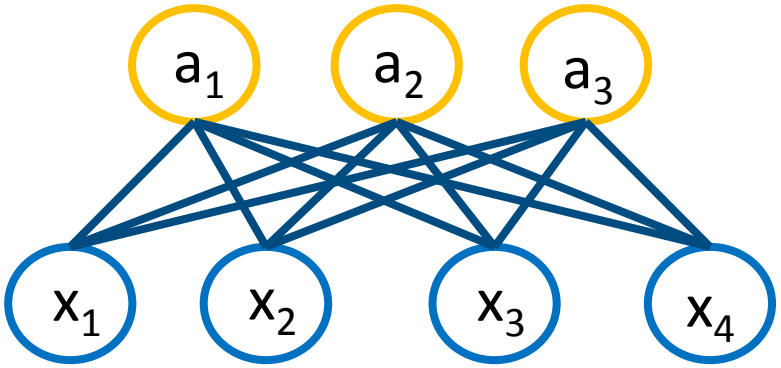
Use Gibbs sampling for inference.

Given observed inputs x, want maximum likelihood estimation:

$$\max_W P(x) = \max_W \sum_a P(x,a)$$

Restricted Boltzmann machine (RBM)

Kay Yu, Andrew Ng
(ECCV 2010)



Layer 2. [a_1, a_2, a_3] (binary-valued)

Input [x_1, x_2, x_3, x_4]

Learning boils down to gradient ascent on P(x)

$$\Delta W_{ij} = \alpha \left([x_i a_j]_{\text{obs}} - [x_i a_j]_{\text{prior}} \right)$$

- $[x_i a_j]_{obs}$ fixing x to observed value, sampling a from $P(a|x)$.
- $[x_i a_j]_{prior}$ from running Gibbs sampling to convergence.

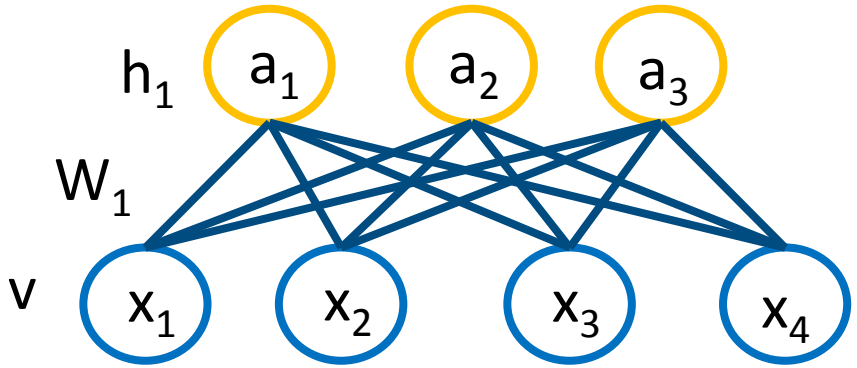
Adding sparsity constraint on the a_i's usually improves results.

DBN Greedy Learning (Hinton et al. 2006)

Honglak Lee
(NIPS 2010)

First step

- Construct an RBM with an input layer *v* and a hidden layer *h*
- Train the RBM



Matteo Matteucci – matteo.matteucci@polimi.it

DBN Greedy Learning (Hinton et al. 2006)

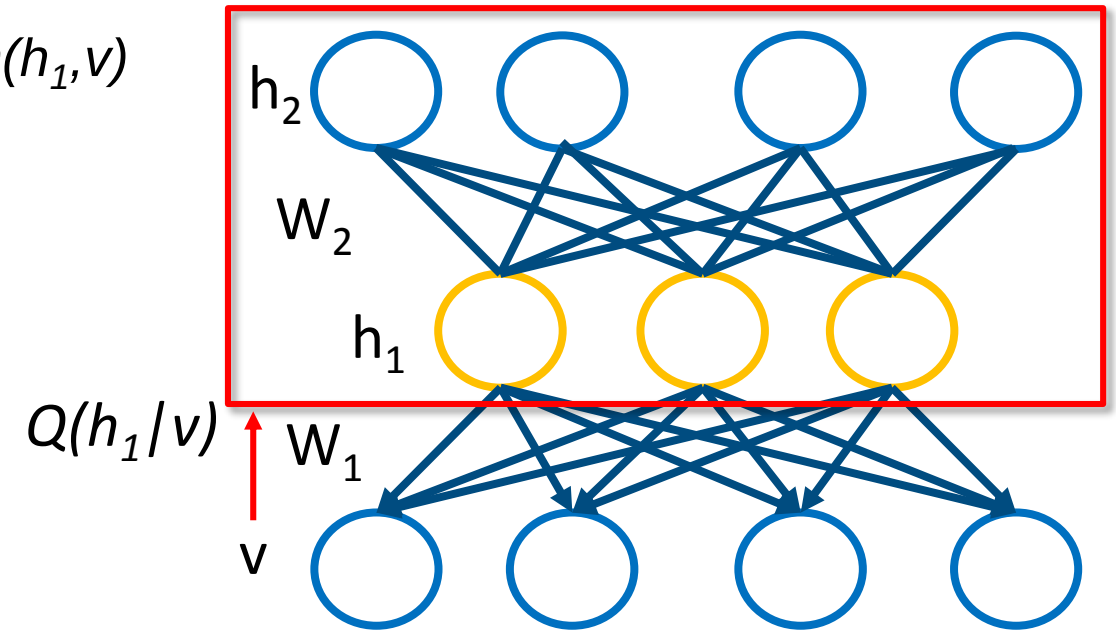
Honglak Lee
(NIPS 2010)

First step

- Construct an RBM with an input layer *v* and a hidden layer *h*
- Train the RBM

Second step

- Stack another hidden layer on top of the RBM to form a new RBM
- Fix W_1, sample h_1 from $Q(h_1,v)$ as input.
- Train W_2 as RBM.



Matteo Matteucci – matteo.matteucci@polimi.it

DBN Greedy Learning (Hinton et al. 2006)

Honglak Lee
(NIPS 2010)

First step

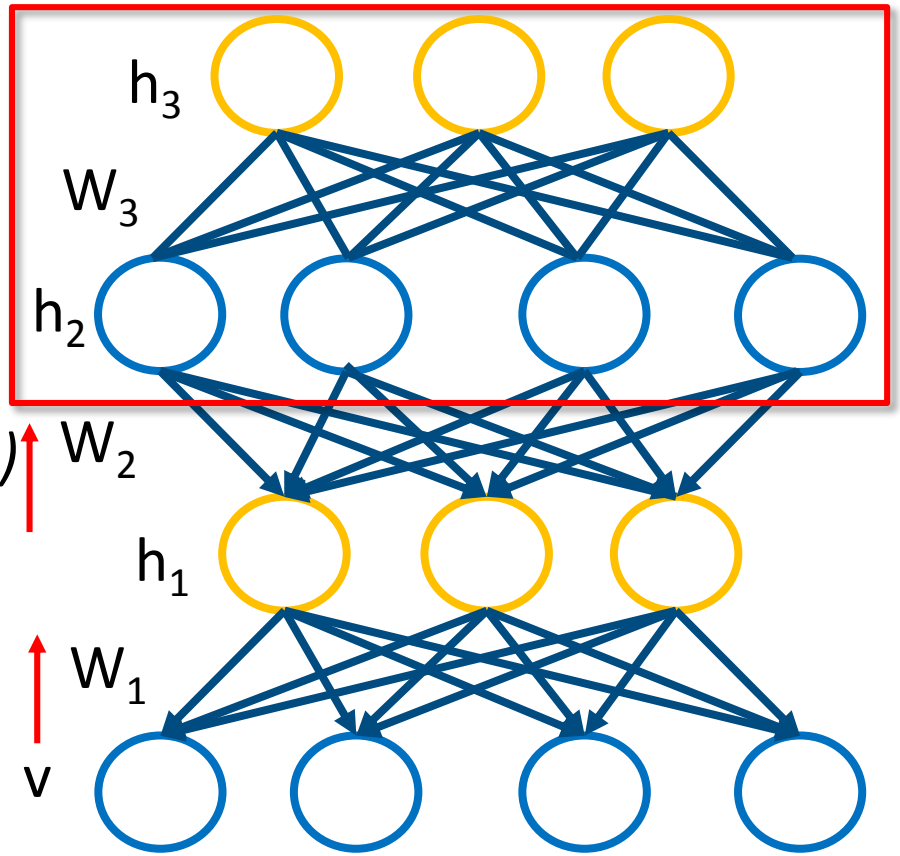
- Construct an RBM with an input layer *v* and a hidden layer *h*
- Train the RBM

Second step

- Stack another hidden layer on top of the RBM to form a new RBM
- Fix W_1, sample h_1 from $Q(h_1|v)$ as input.
- Train W_2 as RBM

Third step

- Continue to stack layers on top of the network, train it as previous step, with samples from $Q(h_1|h_2)$



Matteo Matteucci – matteo.matteucci@polimi.it

DBN Greedy Learning (Hinton et al. 2006)

Honglak Lee
(NIPS 2010)

First step

- Construct an RBM with an input layer *v* and a hidden layer *h*
- Train the RBM

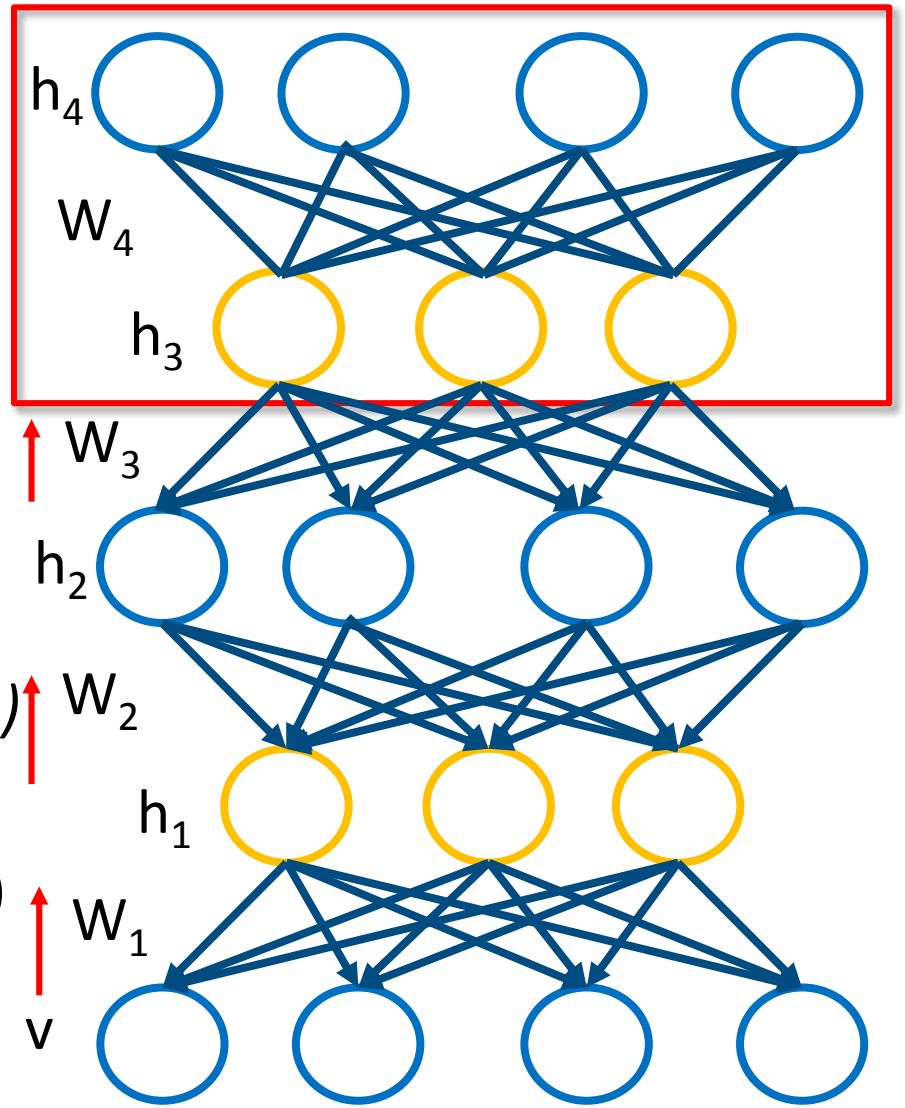
Second step

- Stack another hidden layer on top of the RBM to form a new RBM
- Fix W_1, sample h_1 from $Q(h_1|v)$ as input.
- Train W_2 as RBM

Third step

- Continue to stack layers on top of the network, train it as previous step, with samples from $Q(h_1|h_2)$

And so on …



Matteo Matteucci – matteo.matteucci@polimi.it

How does it work? (Hinton et al. 2006)

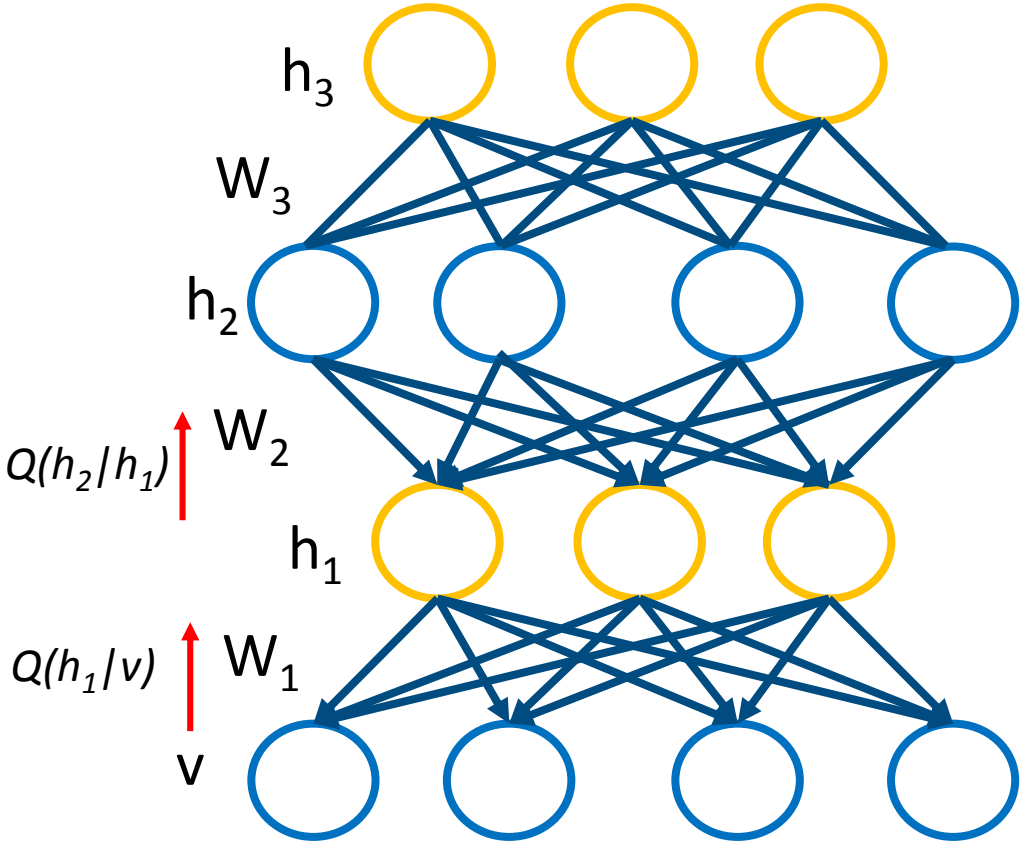
Honglak Lee
(NIPS 2010)

Easy approximate inference:

- $P(h_{k+1}|h_k)$ approximated from the associated RBM
- Approximation because $P(h_{k+1})$ differs between RBM and DBN

Training:

- Variational bound justifies greedy layerwise training of RBMs



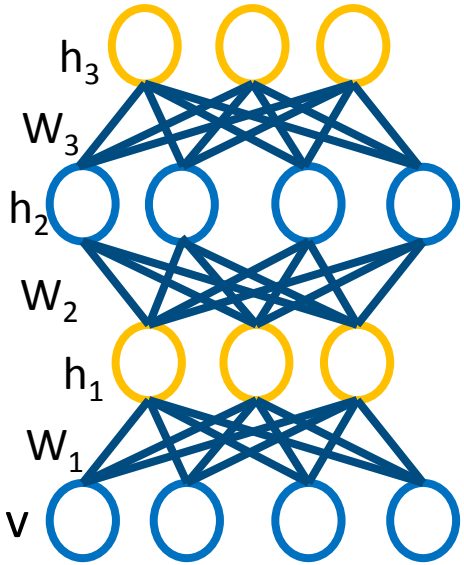
$$\log P(\mathbf{x}) \geq H_{Q(\mathbf{h}|\mathbf{x})} + \sum_{\mathbf{h}} Q(\mathbf{h}|\mathbf{x}) \left(\log P(\mathbf{h}) + \log P(\mathbf{x}|\mathbf{h})\right)$$

Trained by the second layer RBM

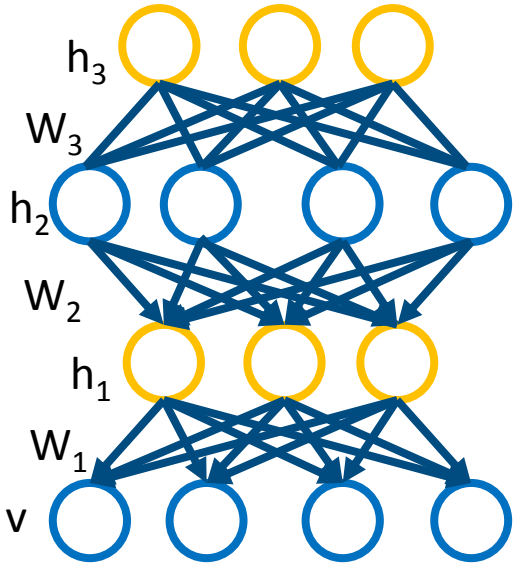
Deep Boltzmann Machines

Honglak Lee
(NIPS 2010)

Undirected connections between all layers



Deep Boltzmann Machine



Deep Belief Network

The training algorithm in Deep Boltzmann Machines has three phases

- Pre-training: initialize the model from stacked RBM
- Generative fine-tuning: variational approximation and persistent chain
- Discriminative fine-tuning: backpropagation w.r.t. output class

POLITECNICO DI MILANO



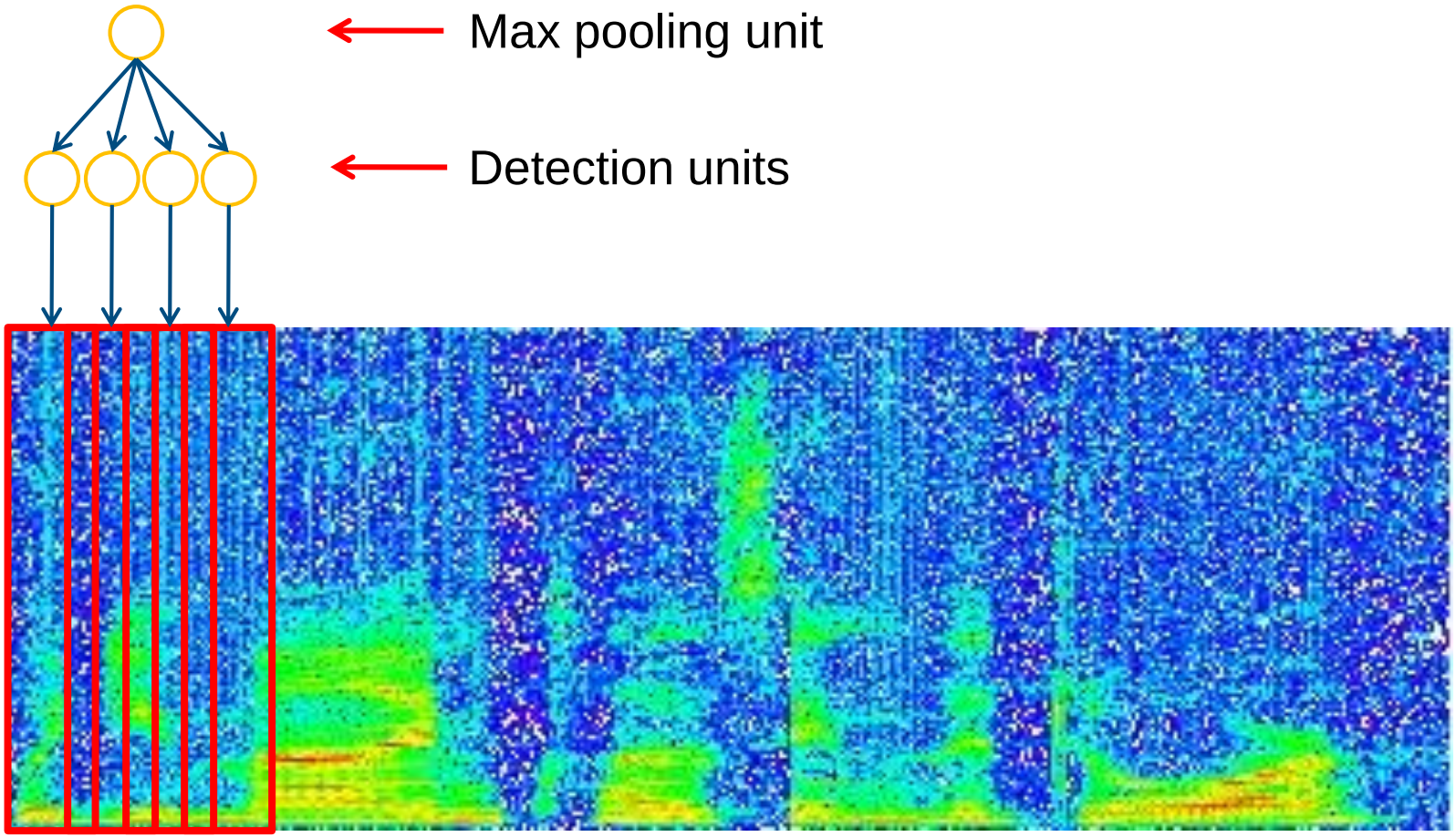


Deep Learning – Convolutional DBN Examples

Matteo Matteucci – matteo.matteucci@polimi.it

Convolutional DBN for audio

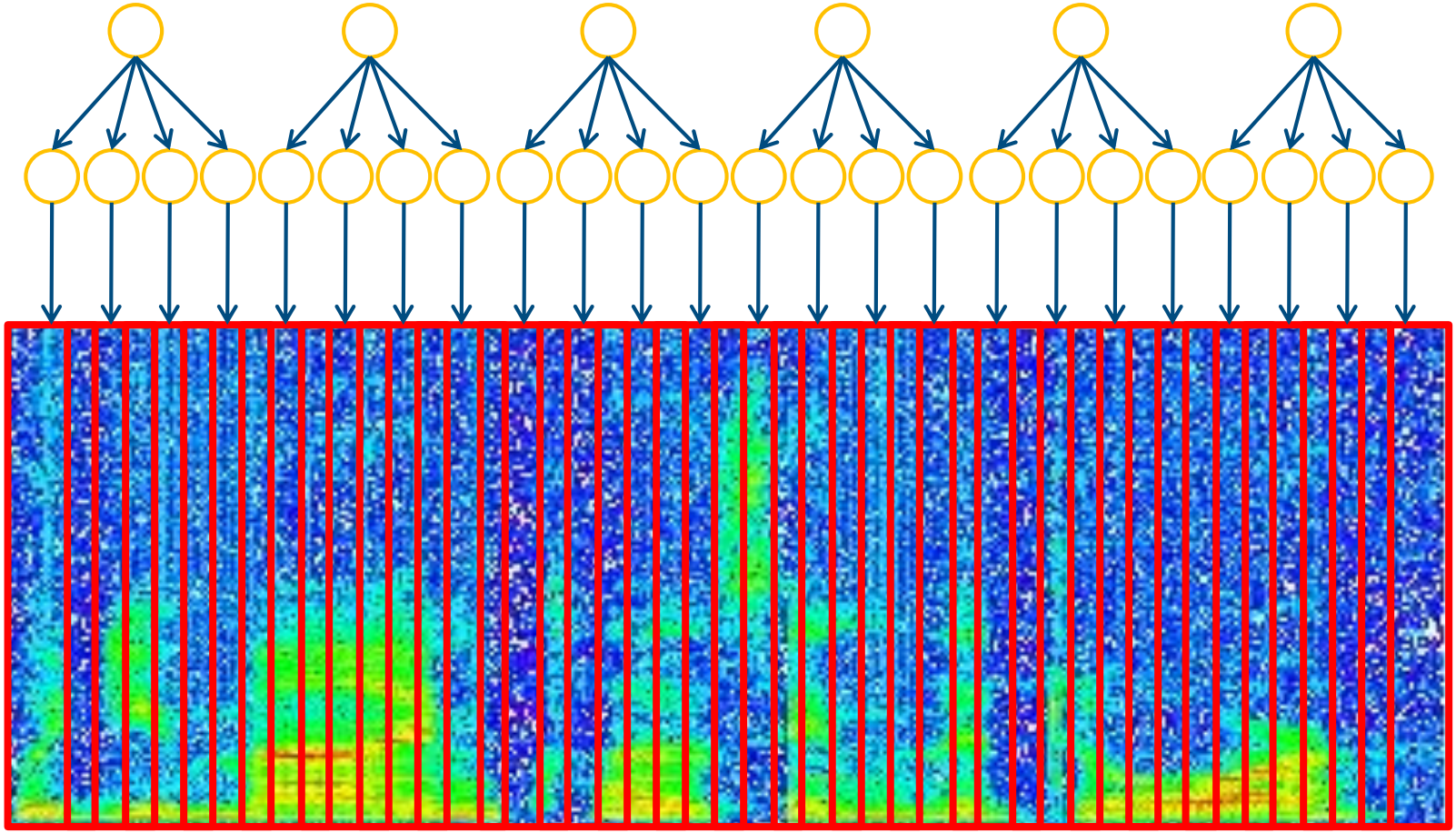
Kay Yu, Andrew Ng
(ECCV 2010)



Spectrogram

Convolutional DBN for audio

Kay Yu, Andrew Ng
(ECCV 2010)



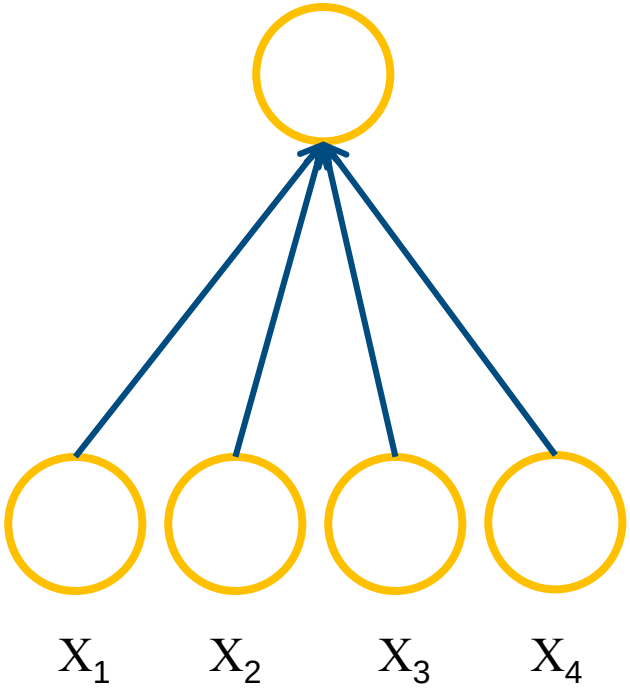
Spectrogram

Probabilistic max pooling

Kay Yu, Andrew Ng
(ECCV 2010)

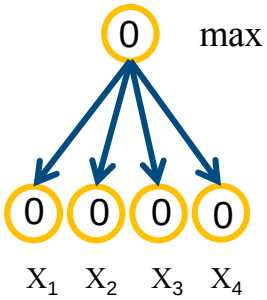
Convolutional NN

$\max \{x_1, x_2, x_3, x_4\}$

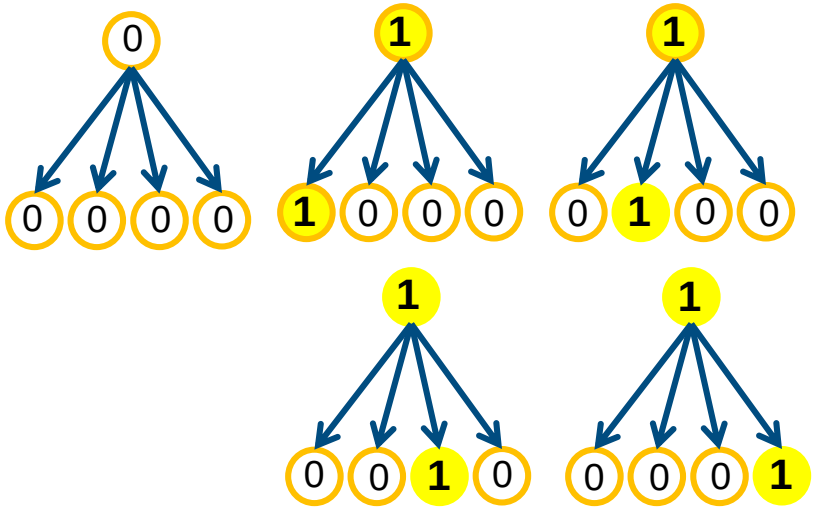


Where x_i are real numbers.

Convolutional DBN



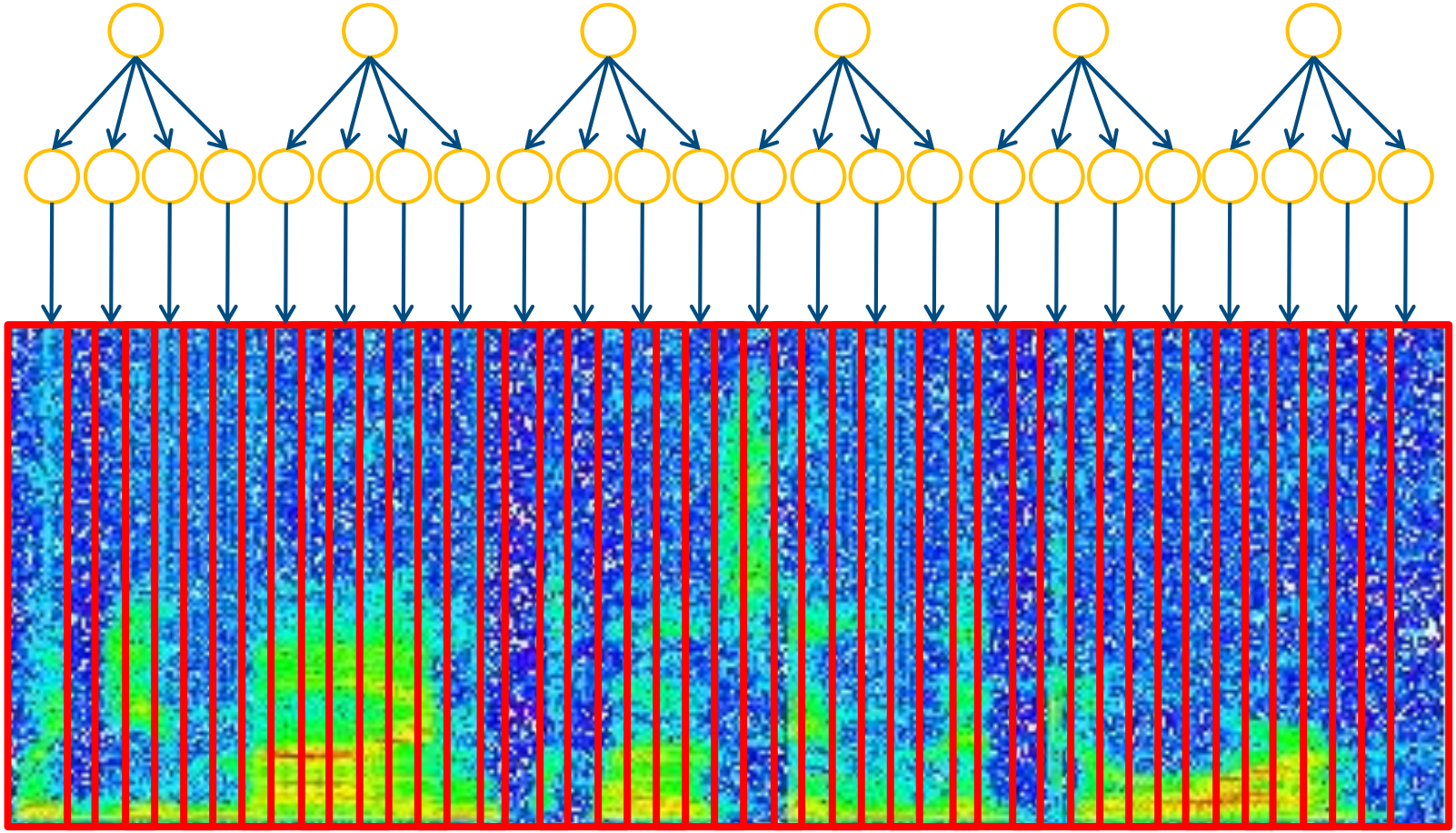
Where x_i are {0,1}, and *mutually exclusive*. Thus, 5 possible cases:



Collapse 2^n configurations into n+1 configurations. Allows inference.

Convolutional DBN for audio

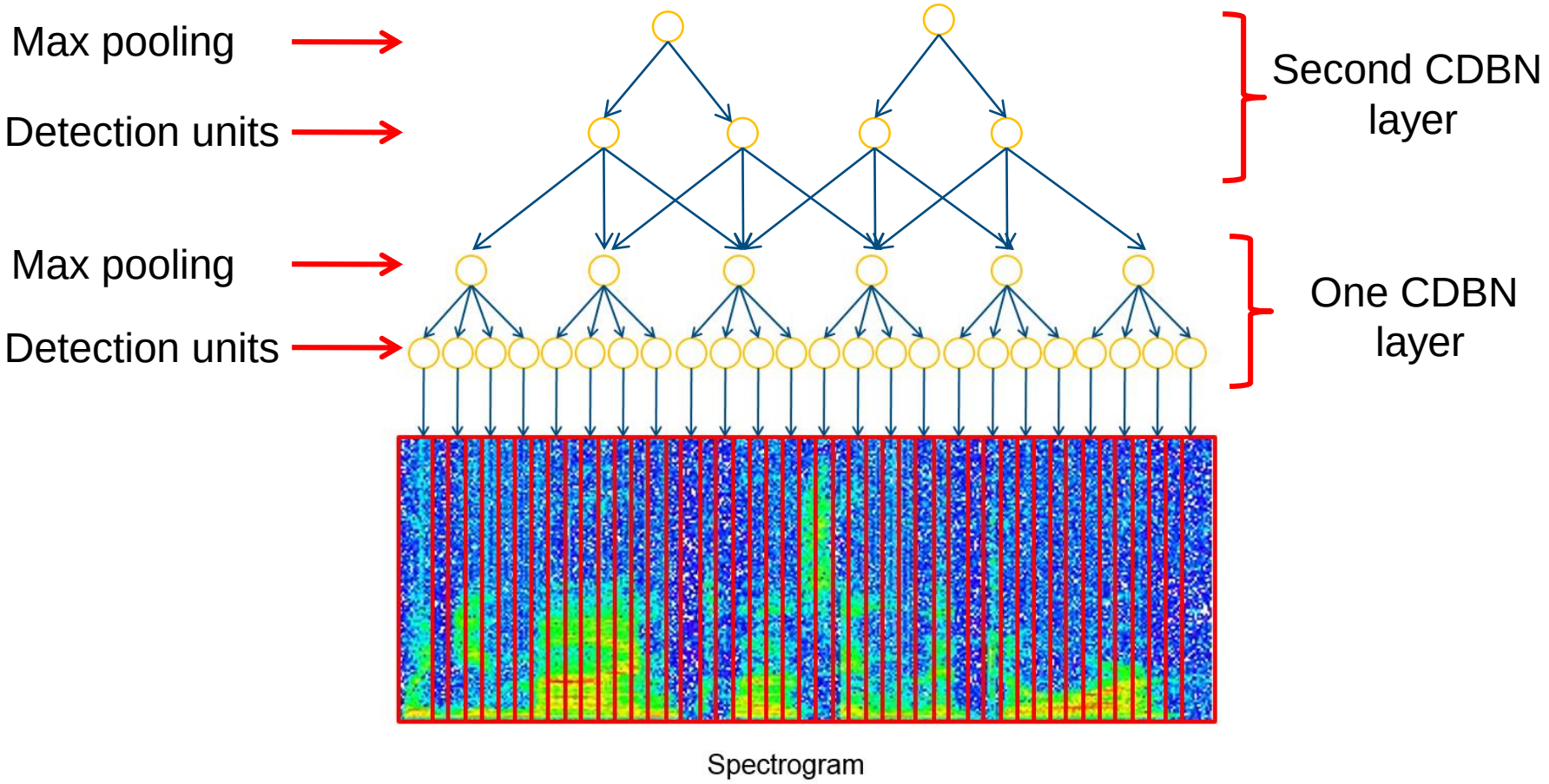
Kay Yu, Andrew Ng
(ECCV 2010)



Spectrogram

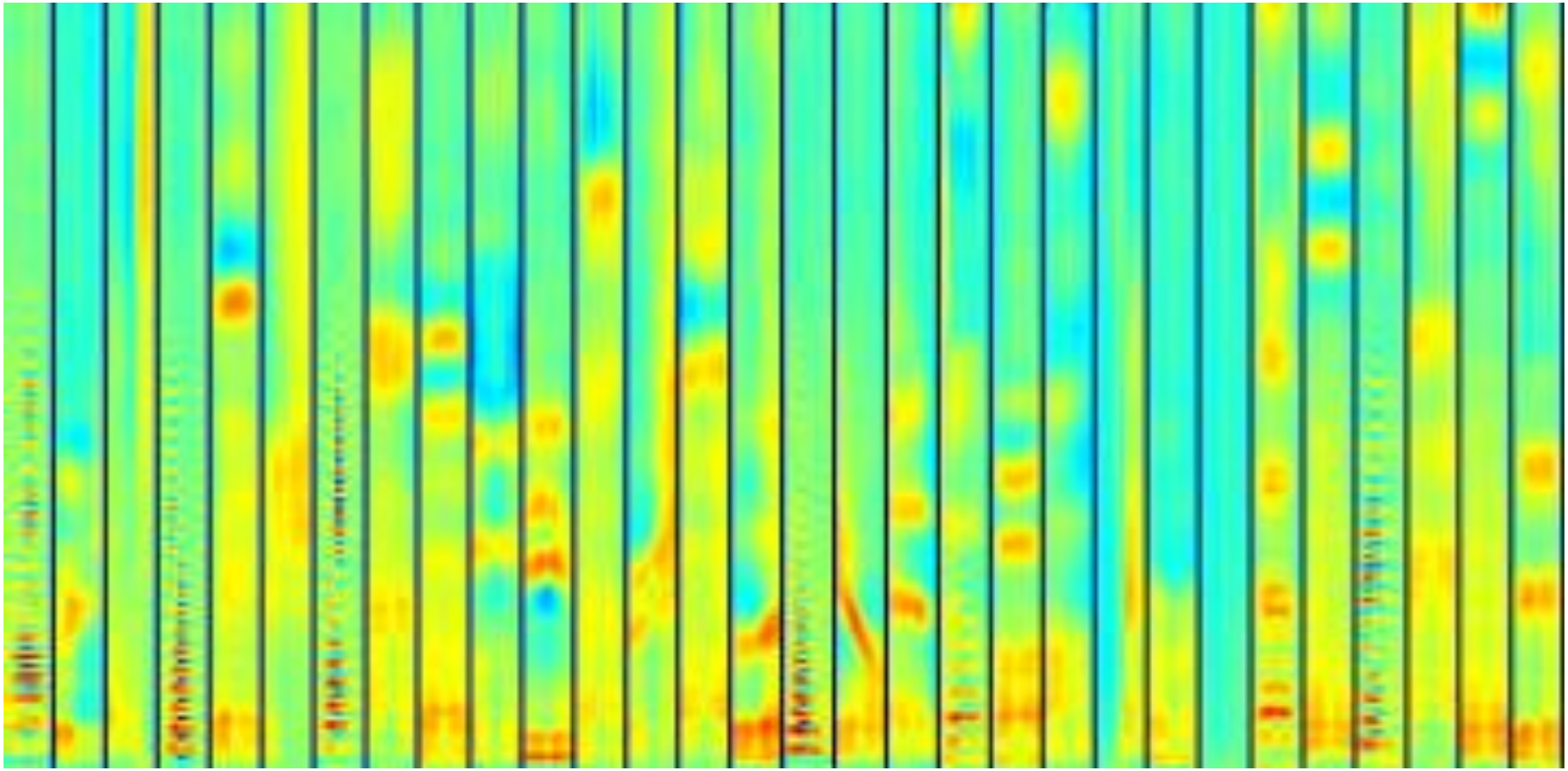
Convolutional DBN for audio

Kay Yu, Andrew Ng
(ECCV 2010)



CDBNs for speech

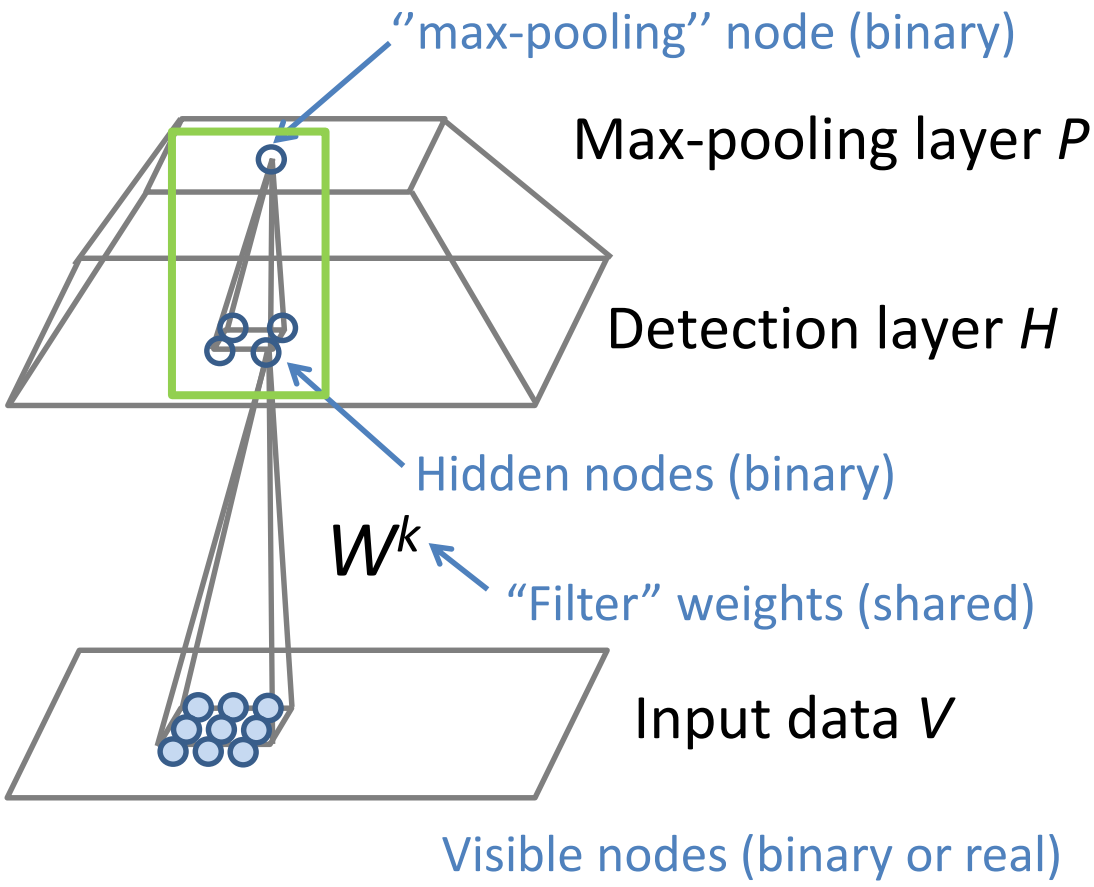
Kay Yu, Andrew Ng
(ECCV 2010)

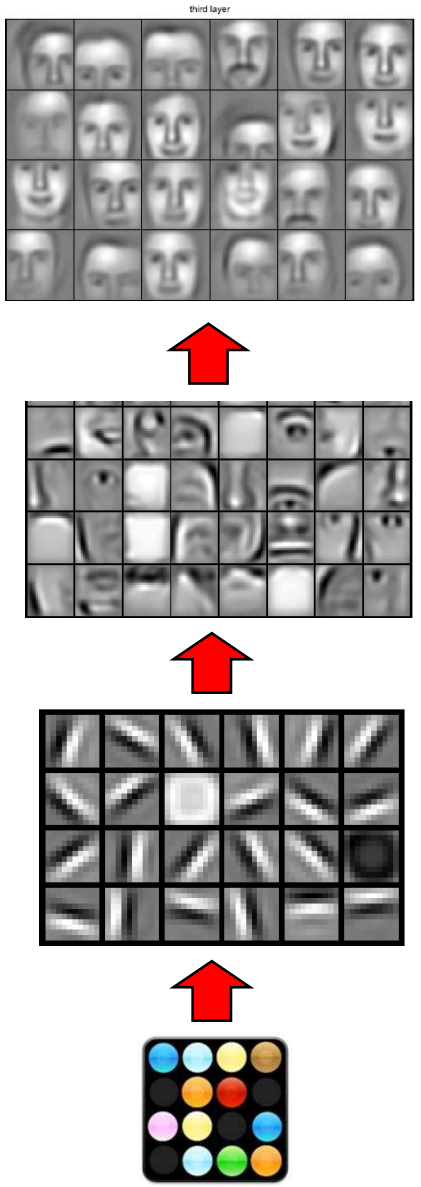


Learned first-layer bases

Convolutional DBN for Images

Kay Yu, Andrew Ng
(ECCV 2010)

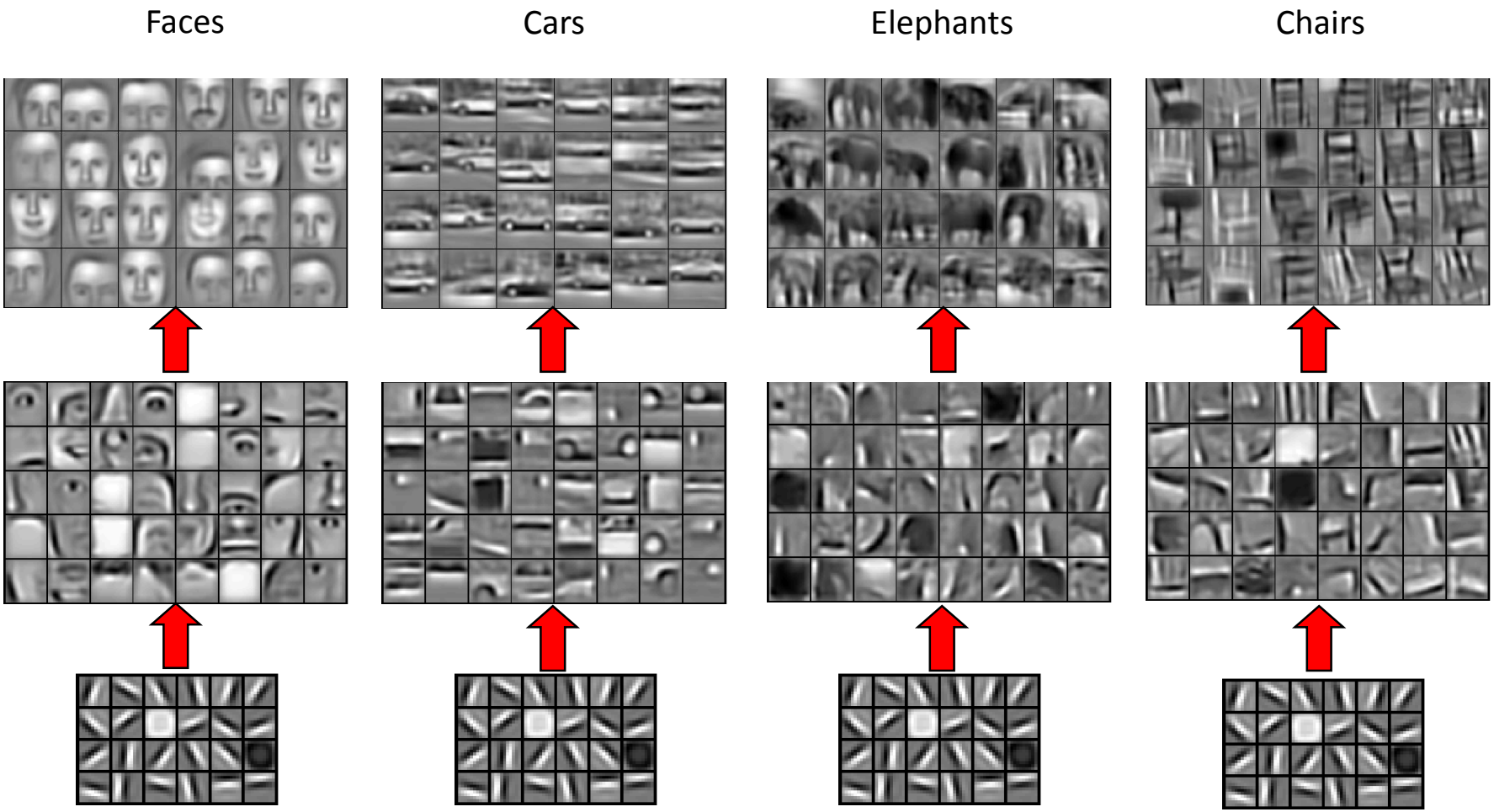




Learning of object parts

Kay Yu, Andrew Ng
(ECCV 2010)

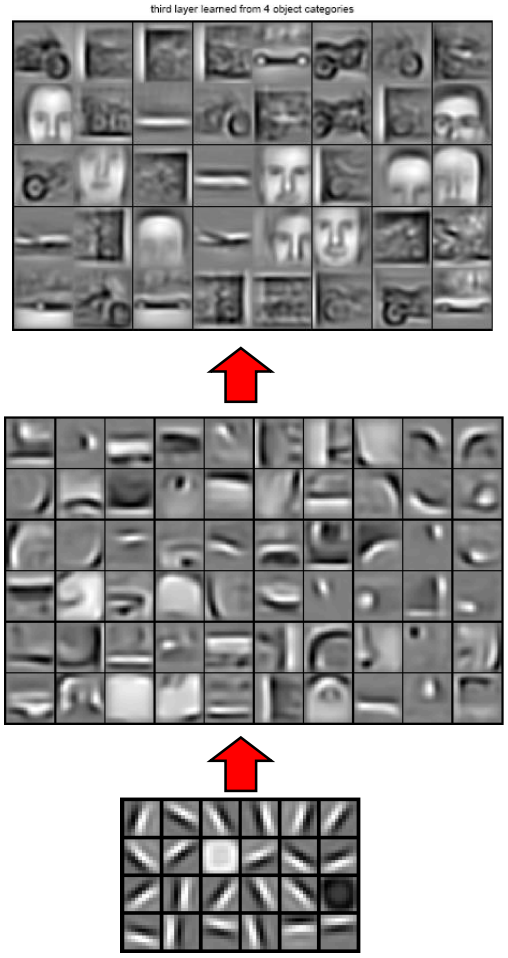
Examples of learned object parts from object categories



Training on multiple objects

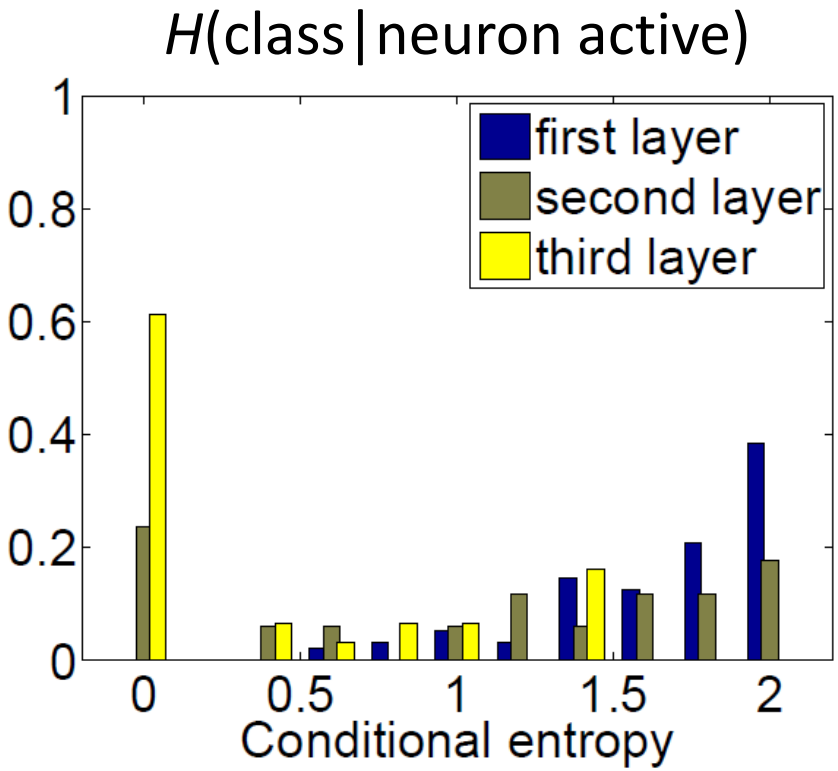
Kay Yu, Andrew Ng
(ECCV 2010)

Trained on 4 classes (cars, faces, motorbikes, airplanes).



More specific features.

Shared-features and object-specific features.



Hierarchical probabilistic inference

Kay Yu, Andrew Ng
(ECCV 2010)

Generating posterior samples from faces by “filling in” experiments (cf. Lee and Mumford, 2003). Combine bottom-up and top-down inference.

Input images

Samples from feedforward Inference (control)

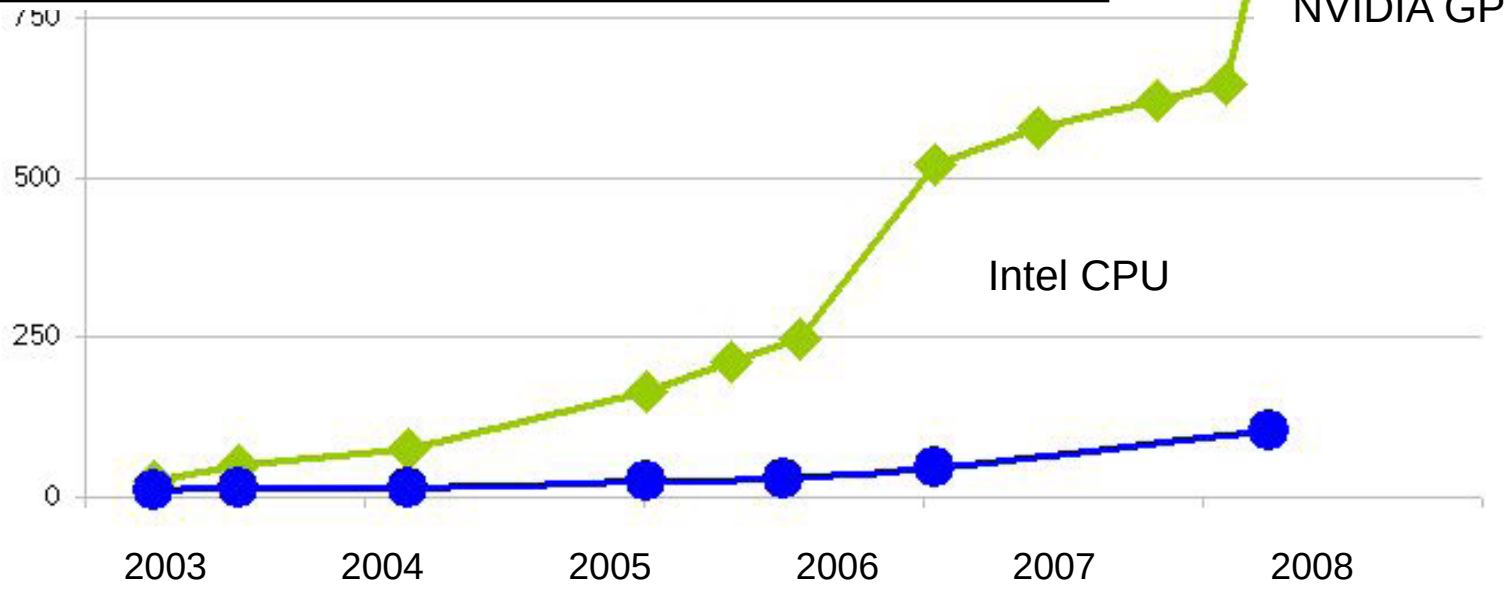
Samples from Full posterior inference

Scaling up with graphics processors

Kay Yu, Andrew Ng
(ECCV 2010)

Published source	Application	Params
Hinton et al., 2006	Digit images	1.6mn
Hinton & Salakhutdinov	Face images	3.8mn
Salakhutdinov & Hinton	Sem. hashing	2.6mn
Ranzato & Szummer	Text	3mn
Using GPU (Raina et al., 2009)		100mn

US$ 250



(Source: NVIDIA CUDA Programming Guide)

Some benchmarks ...

Kay Yu, Andrew Ng
(ECCV 2010)

Audio

TIMIT Phone classification	Accuracy
Prior art (Clarkson et al.,1999)	79.6%
Stanford Feature learning	**80.3%**

TIMIT Speaker identification	Accuracy
Prior art (Reynolds, 1995)	99.7%
Stanford Feature learning	**100.0%**

Images

CIFAR Object classification	Accuracy
Prior art (Yu and Zhang, 2010)	74.5%
Stanford Feature learning	**75.5%**

NORB Object classification	Accuracy
Prior art (Ranzato et al., 2009)	94.4%
Stanford Feature learning	**96.2%**

Video

UCF activity classification	Accuracy
Prior art (Kalser et al., 2008)	86%
Stanford Feature learning	**87%**

Hollywood2 classification	Accuracy
Prior art (Laptev, 2004)	47%
Stanford Feature learning	**50%**

Multimodal (audio/video)

AVLetters Lip reading	Accuracy
Prior art (Zhao et al., 2009)	58.9%
Stanford Feature learning	**63.1%**

POLITECNICO DI MILANO





Deep Learning – ImageNet

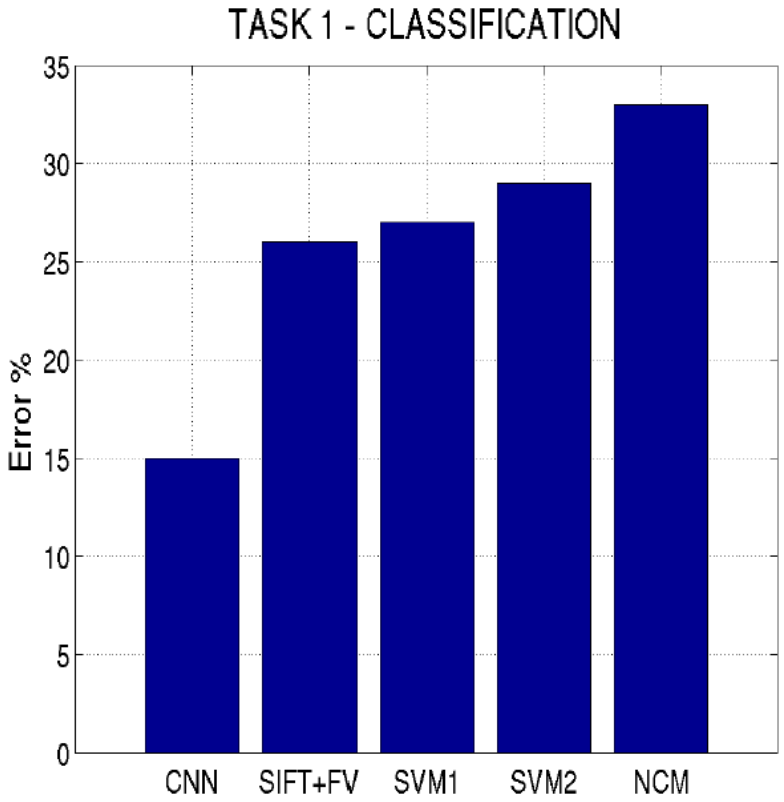
Matteo Matteucci – matteo.matteucci@polimi.it

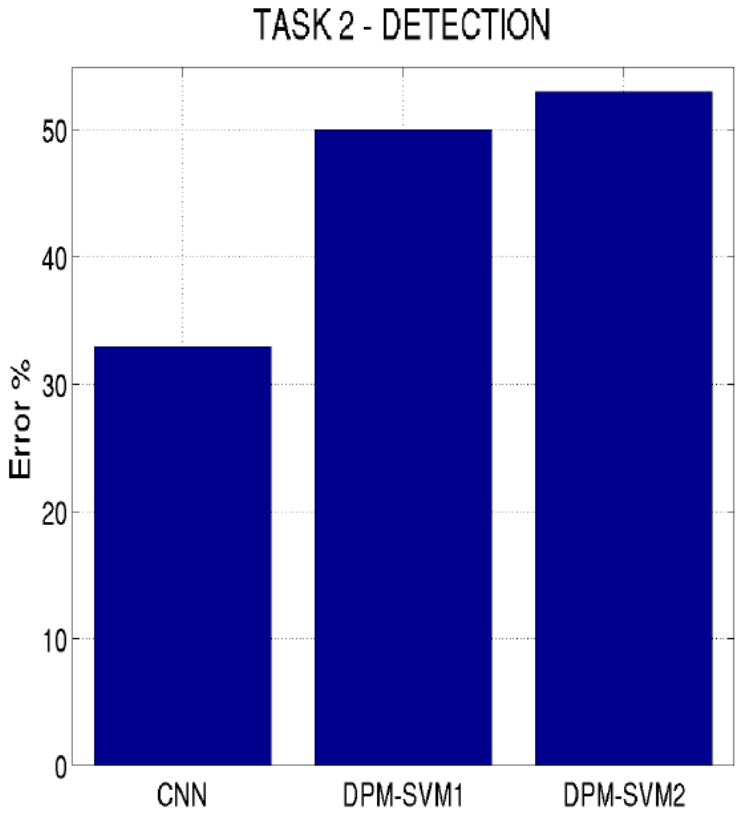
Object Recognition: ILSVRC 2012 Results

Yan LeCun, Marc'Aurelio Ranzato (IMCL 2013)

ImageNet Large Scale Visual Recognition Challenge

- 1000 cathegories
- 1.5 Million labeled training samples





Krizhevsky, Sutskever, Hinton (2012)

Krizhevsky, Sutskever, Hinton (2012)

Yan LeCun, Marc'Aurelio Ranzato (IMCL 2013)

Large convolutional net

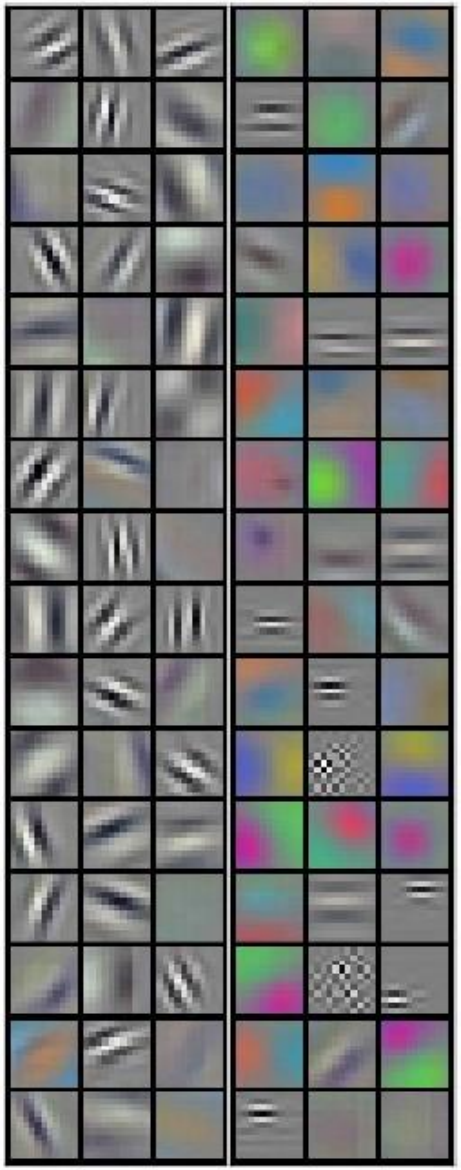
- 650K neurons, 832M synapses, 60M parameters
- Trained with backpropagation on GPU
- Trained «with all the tricks Yann came up with in the last 20 years, plus dropout» (Hinton NIPS'10)
- Image preprocessing: contrast normalization, rectification, etc.

Error rate: 15% (whenever the correct class isn't in top 5)

Previous state of the art: 25% error

A revolution in Computer Vision

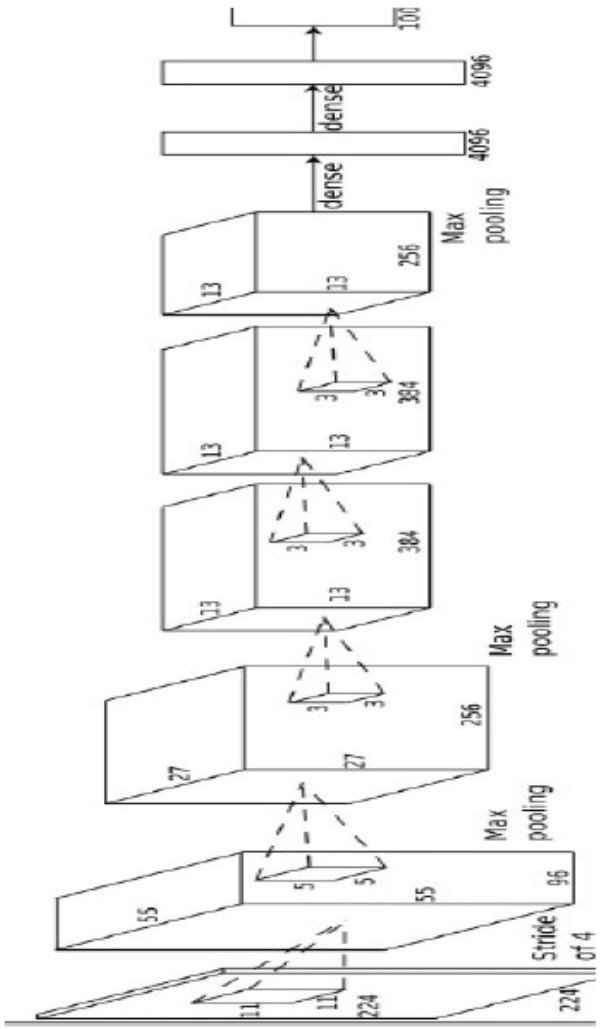
- Acquired by Google in Jan 2013
- Deployed in Google+ Photo Tagging in May 2013



And the winner is ...

Yan LeCun, Marc'Aurelio Ranzato
(IMCL 2013)

4M	FULL CONNECT	4Mflop
16M	FULL 4096/ReLU	16M
37M	FULL 4096/ReLU	37M
	MAX POOLING	
442K	CONV 3x3/ReLU 256fm	74M
1.3M	CONV 3x3ReLU 384fm	224M
884K	CONV 3x3/ReLU 384fm	149M
	MAX POOLING 2x2sub	
	LOCAL CONTRAST NORM	
307K	CONV 11x11/ReLU 256fm	223M
	MAX POOL 2x2sub	
	LOCAL CONTRAST NORM	
35K	CONV 11x11/ReLU 96fm	105M



Matteo Matteucci – matteo.matteucci@polimi.it

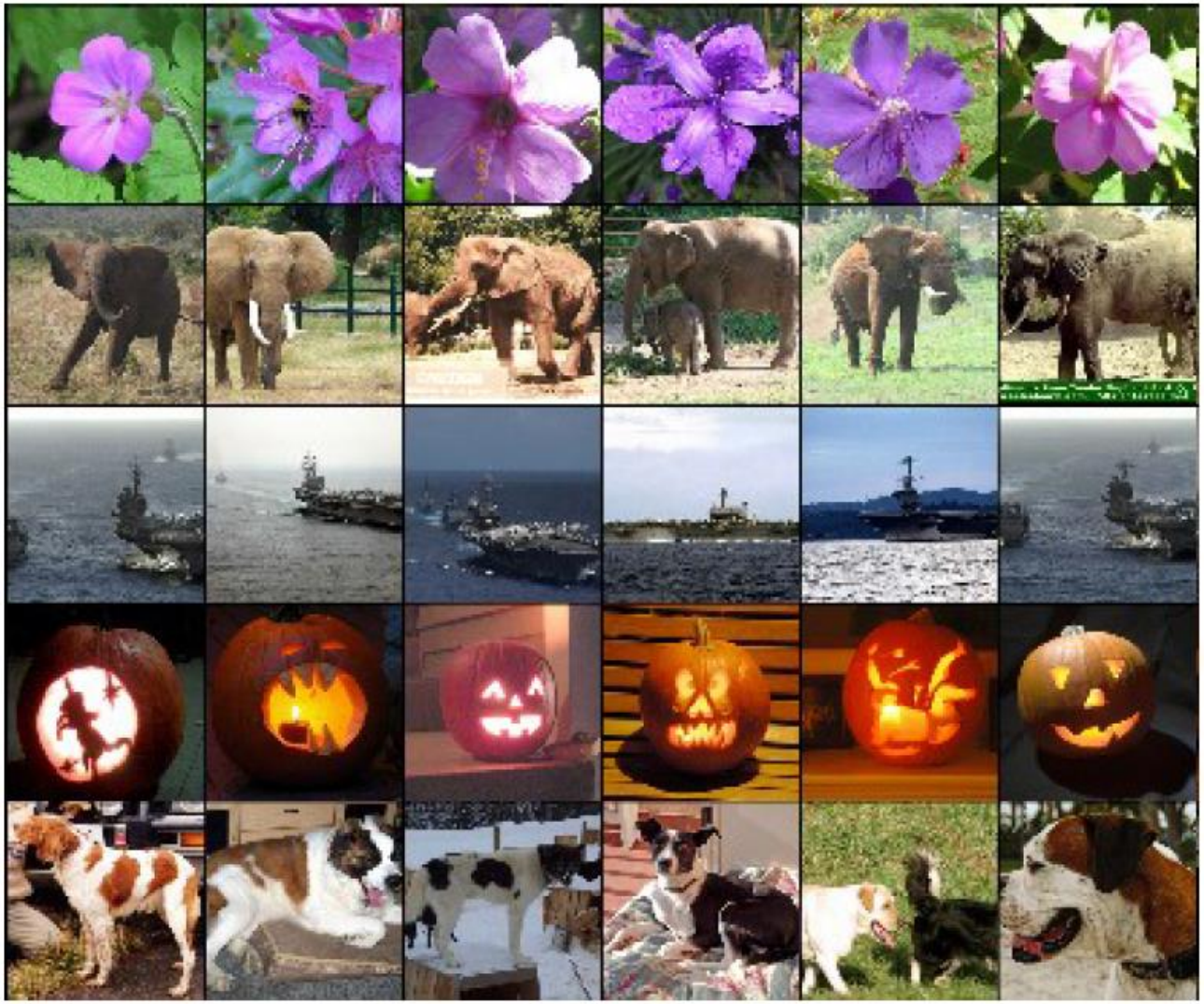
Classification examples ...

Yan LeCun, Marc'Aurelio Ranzato
(IMCL 2013)

Retrieval examples ...

Query

Retrieved Images



Matteo Matteucci – matteo.matteucci@polimi.it

Zeiler and Fergus (2013)

Yan LeCun, Marc'Aurelio Ranzato (IMCL 2013)

Convolutional network

- 8 layers, input 224x224 pixels
- Conv – pool – conv – pool – conv conv – conv – full – full – full
- Rectified-linear Units (ReLU)
- Divisive contrast normalization across features [Jarret et al. 2009]

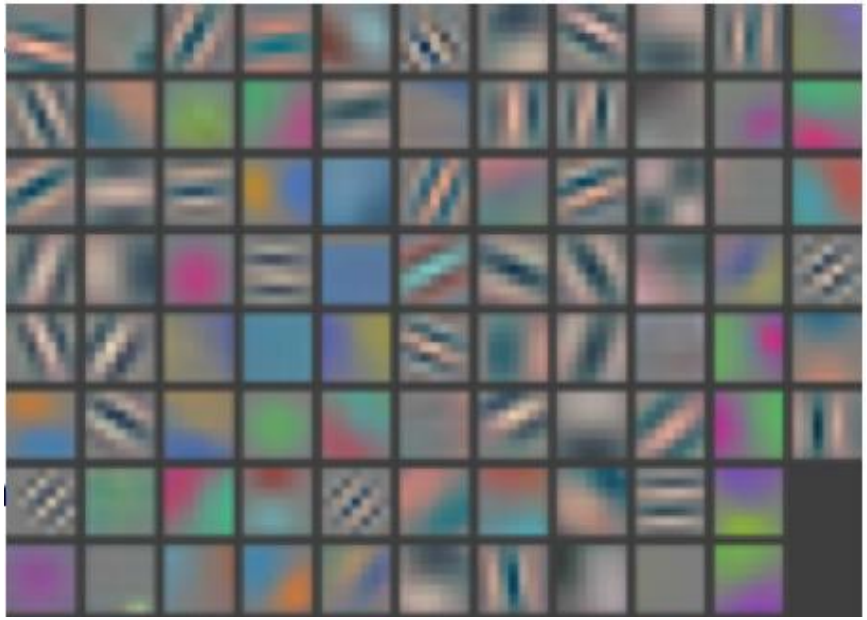
Trained on ImageNet 2012 training set

- 1.3M images, 1000 classes
- 10 different crops/flips per image

Stochastic gradient descent

- 70 epochs (7-10 days)
- Learning rate anealing

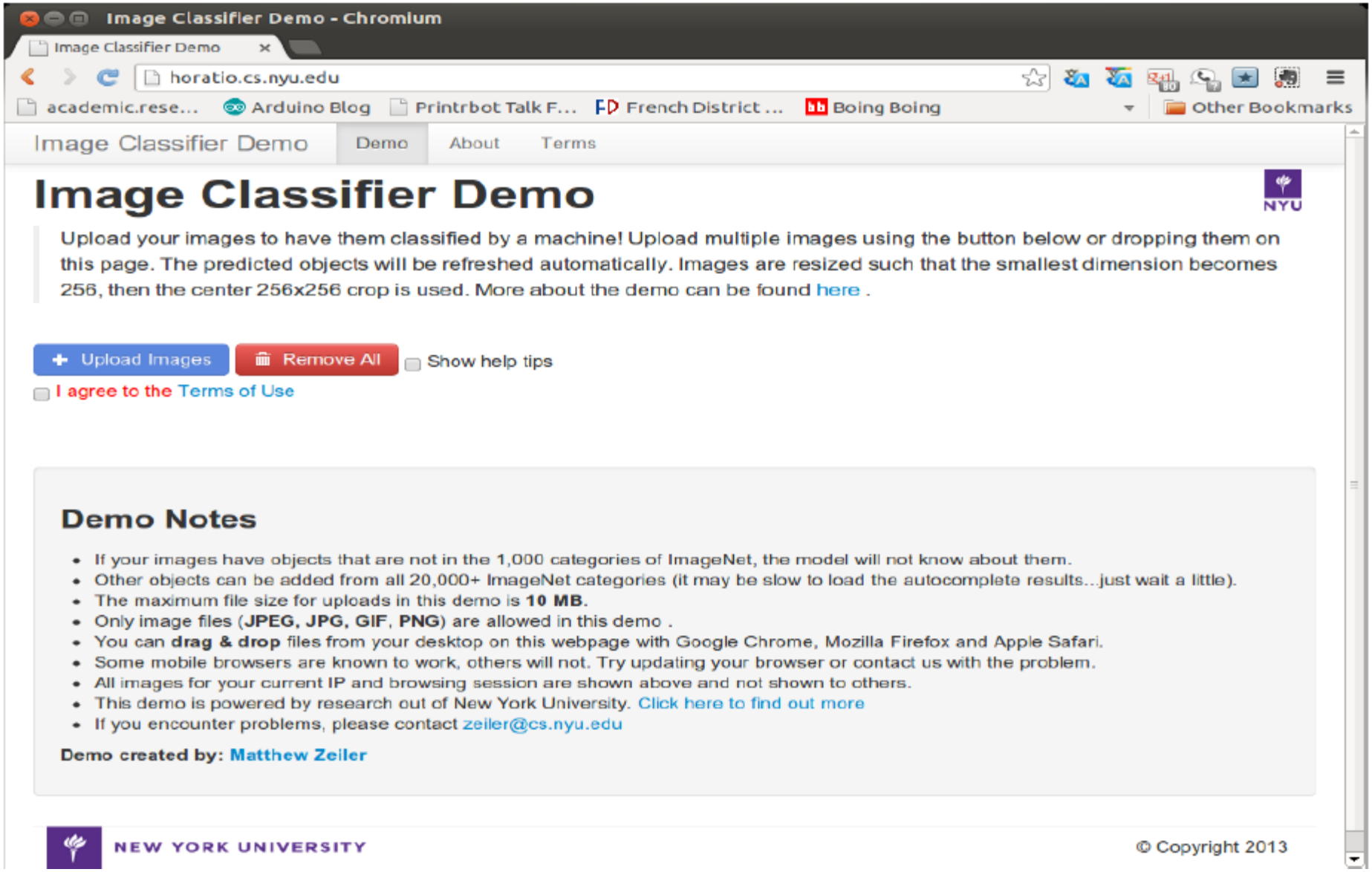
Regularization with dropout



Clarifai	11.7%	Deep CNN
NUS	13.0%	SVM based + Deep CNN
ZF	13.5%	Deep CNN
Andrew Howard	13.6%	Deep CNN
OverFeat-NYU	14.1%	Deep CNN
UvA-Euvison	14.2%	Deep CNN

http://horatio.cs.nyu.edu/

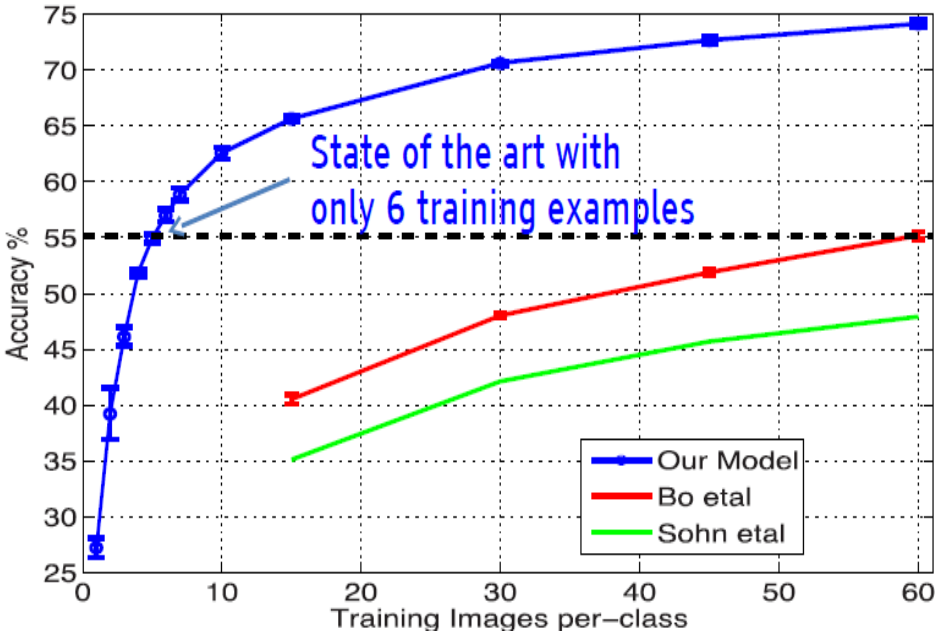
Yan LeCun, Marc'Aurelio Ranzato
(IMCL 2013)



Features are generic

Yan LeCun, Marc'Aurelio Ranzato (IMCL 2013)

- Network trained on ImageNet first
- Last layer chopped off
- Last layer trained on Caltech 256 keeping previous layers fixed
- State of the art accuracy with only 6 samples/class



- Similar results obtained on PASCAL VOC 2012 image recognition dataset

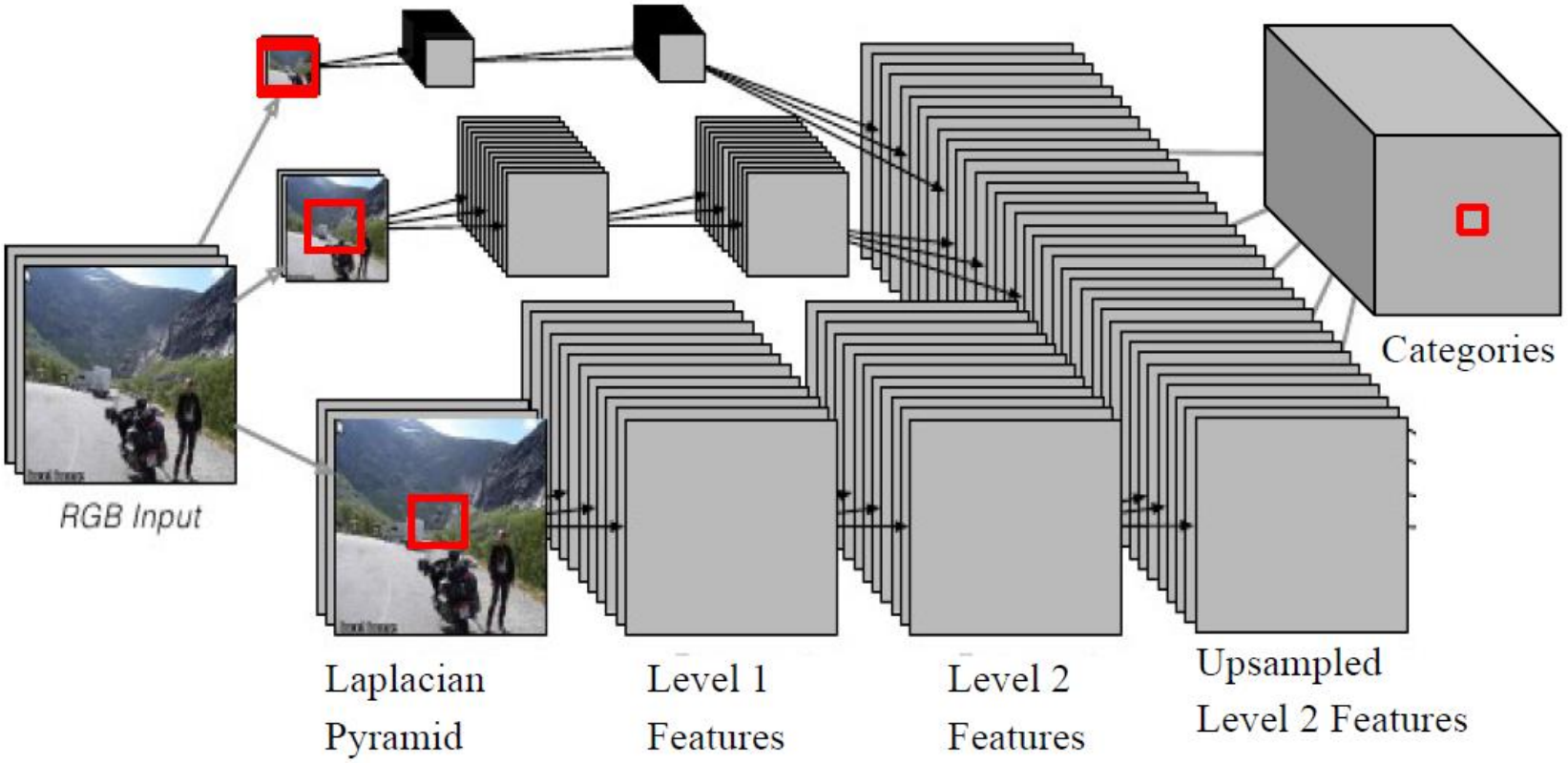
Acc %	[15]	[19]	Ours	Acc %	[15]	[19]	Ours
Airplane	92.0	**97.3**	96.0	Dining table	63.2	**77.8**	67.7
Bicycle	74.2	**84.2**	77.1	Dog	68.9	83.0	**87.8**
Bird	73.0	80.8	**88.4**	Horse	78.2	**87.5**	86.0
Boat	77.5	85.3	**85.5**	Motorbike	81.0	**90.1**	85.1
Bottle	54.3	**60.8**	55.8	Person	91.6	**95.0**	90.9
Bus	85.2	**89.9**	85.8	Potted plant	55.9	**57.8**	52.2
Car	81.9	**86.8**	78.6	Sheep	69.4	79.2	**83.6**
Cat	76.4	89.3	**91.2**	Sofa	65.4	**73.4**	61.1
Chair	65.2	**75.4**	65.0	Train	86.7	**94.5**	91.8
Cow	63.2	**77.8**	74.4	Tv/monitor	77.4	**80.7**	76.1
Mean	74.3	**82.2**	79.0	# won	0	**15**	5

ConvNet for scene parsing

Yan LeCun, Marc'Aurelio Ranzato
(IMCL 2013)

Each output sees a large input context

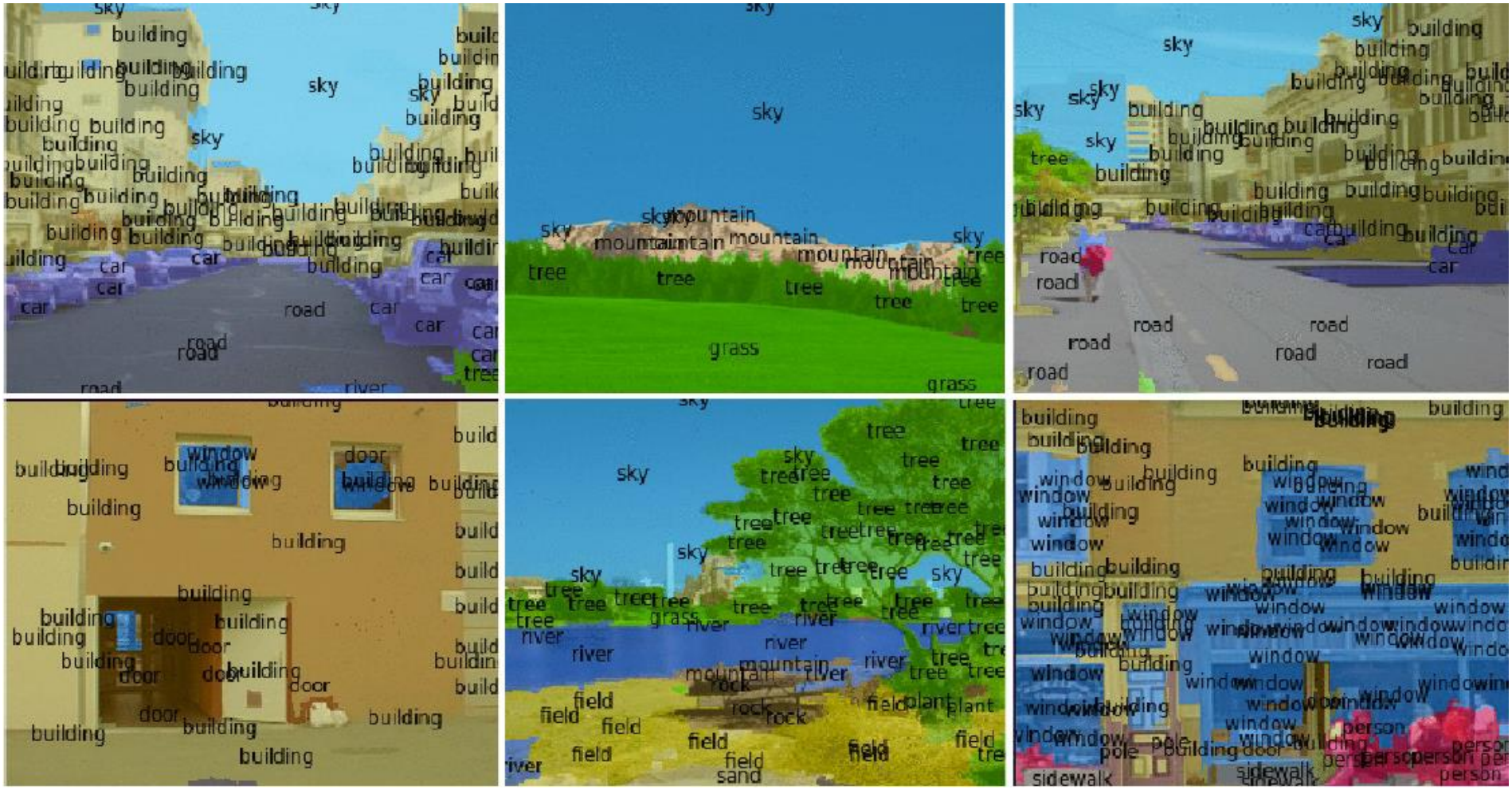
- 46x46 window at full res., 92x92 and ½ res., 184x814 at ¼ res.
- [7x7conv] →[2x2pool]→[7x7conv] →[2x2pool] →[7x7conv] →
- Trained supervised on fully labeled images



Semantic labeling

Yan LeCun, Marc'Aurelio Ranzato
(IMCL 2013)

Labeling each pixel with the object it belongs to ...



[Farabet et al. ICML 2012, PAMI 2013]

Outdoor scene interpretation

Yan LeCun, Marc'Aurelio Ranzato
(IMCL 2013)

No post processing required

Frame by frame scene interpretation

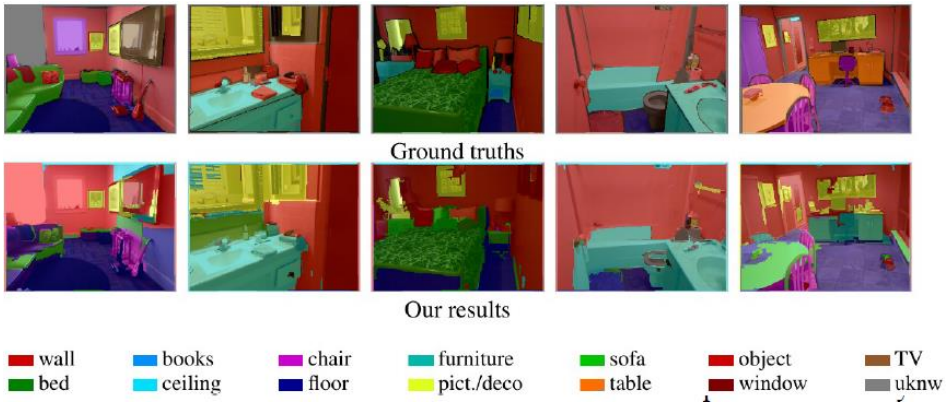
Runs at 50ms/frame on Virtes-6 FPGA Hardware

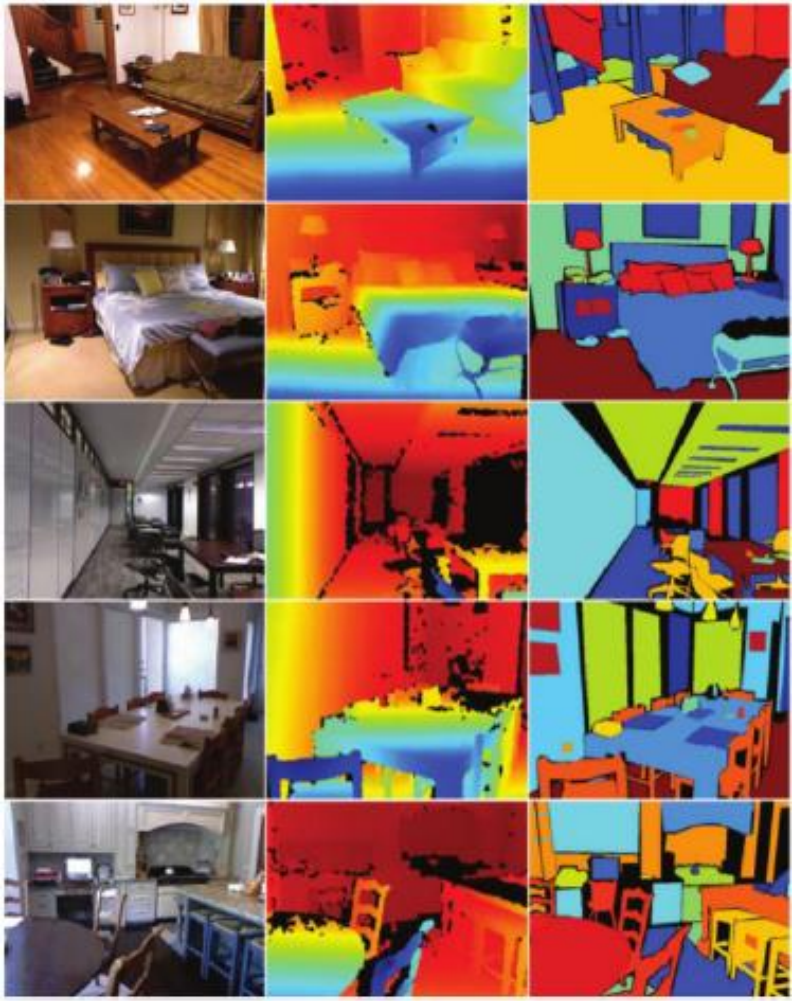
Indoor scene interpretation

Yan LeCun, Marc'Aurelio Ranzato
(IMCL 2013)

NYU RGB-Deph Indoor Dataset

- 407024 RGBD images
- 1449 labeled frames
- 849 object categories





POLITECNICO DI MILANO





Deep Learning – Deep Questions?

Matteo Matteucci – matteo.matteucci@polimi.it