香港中文大學
The Chinese University of Hong Kong

ELEG 5491
Introduction to Deep Learning

Xiaogang Wang

xgwang@ee.cuhk.edu.hk

Department of Electronic Engineering,
The Chinese University of Hong Kong

Course Information

- Course webpage

http://www.ee.cuhk.edu.hk/~xgwang/dl/

- Discussions
  - WeChat account @DeepLearningCUHK
  - Twitter account @dl_cuhk
  - WeChat group (see QR code on webpage)
  - Notes at Github (https://eleg5491.github.io/)

Course Information

- Instructor: Xiaogang Wang
  - SHB 415
  - Office hours: after Tuesday's class or by appointment
- Tutor: Hongyang Li (leader)
  - SHB 301
  - yangli@ee.cuhk.edu.hk
  - Office hours: 10:00 – 12:00 on Wednesday

Course Information

- Tutor: Tong Xiao
  - SHB 304
  - xiaotong@ee.cuhk.edu.hk
  - Office hours: 14:40-16:30 on Monday
- Tutor: Wei Yang
  - SHB 304
  - wyang@ee.cuhk.edu.hk
  - Office hour: 9:30-11:30 on Friday

Course Information

- Lecture time & venue
  - Tuesday: 14:30 – 15:15, LT, Basic Medicine Science Building
  - Thursday: 14:30 – 16:15, L4, Science Center
- Unofficial optional tutorials (10 times, one hour each time)
  - Tuesday 15:30 – 16:30
  - Wednesday 16:30 – 17:30
  - Friday 16:30 – 17:30

Course Information

- Homework (30%)
- Quiz 1 (15%)
- Quiz 2 (15%)
- Project (40%)
  - Topics
    - Applications of deep learning
    - Implementation of deep learning
    - Study deep learning algorithms
  - You should submit
    - One page proposal and discuss it with tutor (topic, idea, method, experiments)
    - A term paper of 4 pages (excluding figures) in maximum, double column, font size is equal or larger than 10.
    - Code and sample data
    - Project presentation
    - Poster presentation + tea party
  - No survey
  - No collaboration
  - We can reimburse cloud computing service at Amazon up to 20 hours each person

Course Information

- Examples of project topics
  - Implement CNN with GPU and compare its efficiency with Caffe
  - Fast CPU implementation of CNN
  - We provide a baseline model of GoogLeNet on ImageNet, and you try to improve it
  - Choose one of the deep learning related competitions (such as ImageNet), and compare your result with published ones
  - Propose a deep model to effectively learn dynamic features from videos
  - Deep learning for speech recognition
  - Deep learning for object detection

Textbook

- Ian Goodfellow and Yoshua Bengio and Aaron Courville, “Deep Learning,” MIT Press, 2016

Lectures

Week	Topics	Requirements
1 (Jan 10 & 12)	Introduction	
2 (Jan 17 & 19)	Machine learning basics	
3 (Jan 24 & 26)	Multilayer neural networks	Homework 1
Chinese New Year		
4 (Feb 7 & 9)	Convolutional neural netowrks	Homework 2
5 (Feb 14 & 16)	Optimization for training deep neural networks	
6 (Feb 21 & 23)	Network structures	Quiz 1 (Feb 21)
7 (Feb 28 & Mar 2)	Recurrent neural network (RNN) and LSTM	
8 (Mar 7 & 9)	Deep belief net and auto-encoder	Homework 3
9 (Mar 14 & 16)	Reinforcement learning & deep learning	Project proposal
10 (Mar 21 & 23)	Attention models	
11 (Mar 28 & 30)	Generative adversarial networks (GAN)	
12 (Apr 6)	Structured deep learning	
13 (Apr 11 & 18)	Course sum-up	Quiz 2 (Apr 18)
Project presentation (to be decided)		

Tutorials

Times	Topic
1	Python/Numpy tutorial/AWS tutorial
2	Understand backpropagation
3	Torch tutorial
4	Caffe/Tensorflow/Theano
5	Roadmaps of deep learning models
6	Hands on experiment with debugging models
7	GPU parallel programming
8	Final project proposal discussion
9	Assignment and quiz review
10	Fancy stuff: deep learning on spark, future directions

Hands-on assignments are provided in tutorials. Bring your laptop

Introduction to Deep Learning

Outline

- Historical review of deep learning
- Understand deep learning
- Interpret neural semantics

Machine Learning

$$x \longrightarrow \boxed{F(x)} \longrightarrow y \begin{cases} \text{Class label (Classification)} \\ \text{Vector (Estimation)} \end{cases}$$

Object recognition $\longrightarrow$ {dog, cat, horse, flower, …}

Low-resolution image

Super resolution $\longrightarrow$

High-resolution image

Lectures

Week	Topics	Requirements
1 (Jan 10 & 12)	Introduction	
2 (Jan 17 & 19)	**Machine learning basics**	
3 (Jan 24 & 26)	Multilayer neural networks	Homework 1
	Chinese New Year	
4 (Feb 7 & 9)	Convolutional neural netowrks	Homework 2
5 (Feb 14 & 16)	Optimization for training deep neural networks	
6 (Feb 21 & 23)	Network structures	Quiz 1 (Feb 21)
7 (Feb 28 & Mar 2)	Recurrent neural network (RNN) and LSTM	
8 (Mar 7 & 9)	Deep belief net and auto-encoder	Homework 3
9 (Mar 14 & 16)	Reinforcement learning & deep learning	Project proposal
10 (Mar 21 & 23)	Attention models	
11 (Mar 28 & 30)	Generative adversarial networks (GAN)	
12 (Apr 4 & 6)	Structured deep learning	Quiz 2 (Apr 4)
13 (Apr 11 & 18)	Course sum-up	
	Project presentation (to be decided)	

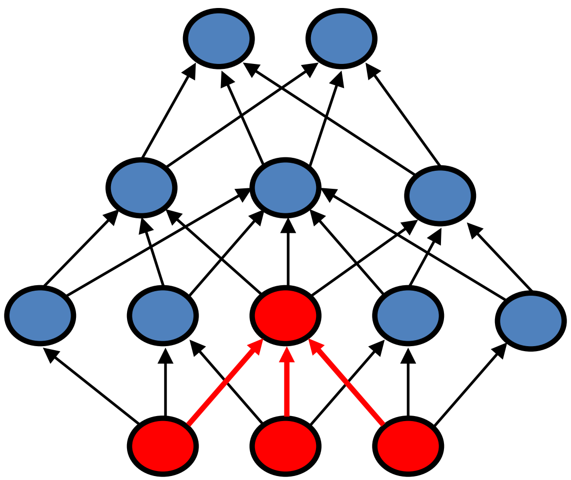
Neural network

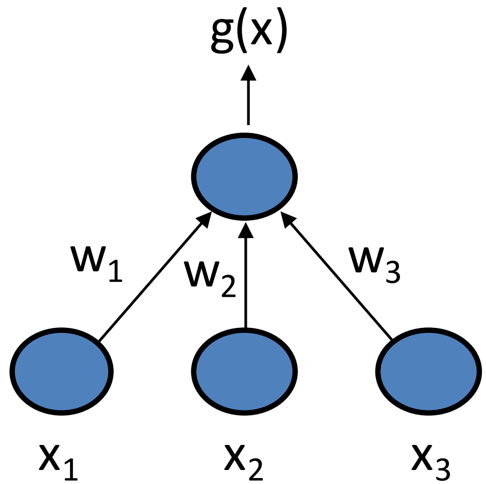
Back propagation

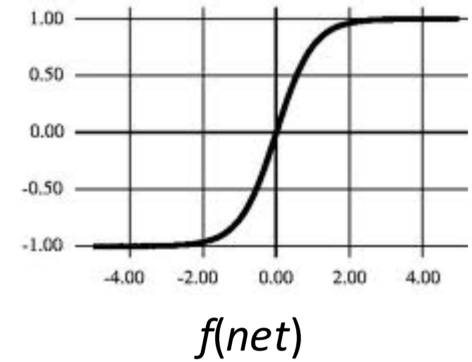
Nature

1940s

1986







$$g(\mathbf{x}) = f(\sum_{i=1}^{d} x_i w_i + w_0) = f(\mathbf{w}^t \mathbf{x})$$

Lectures

Week	Topics	Requirements
1 (Jan 10 & 12)	Introduction	
2 (Jan 17 & 19)	Machine learning basics	
3 (Jan 24 & 26)	**Multilayer neural networks**	**Homework 1**
	Chinese New Year	
4 (Feb 7 & 9)	Convolutional neural netowrks	Homework 2
5 (Feb 14 & 16)	Optimization for training deep neural networks	
6 (Feb 21 & 23)	Network structures	Quiz 1 (Feb 21)
7 (Feb 28 & Mar 2)	Recurrent neural network (RNN) and LSTM	
8 (Mar 7 & 9)	Deep belief net and auto-encoder	Homework 3
9 (Mar 14 & 16)	Reinforcement learning & deep learning	Project proposal
10 (Mar 21 & 23)	Attention models	
11 (Mar 28 & 30)	Generative adversarial networks (GAN)	
12 (Apr 4 & 6)	Structured deep learning	Quiz 2 (Apr 4)
13 (Apr 11 & 18)	Course sum-up	
	Project presentation (to be decided)	

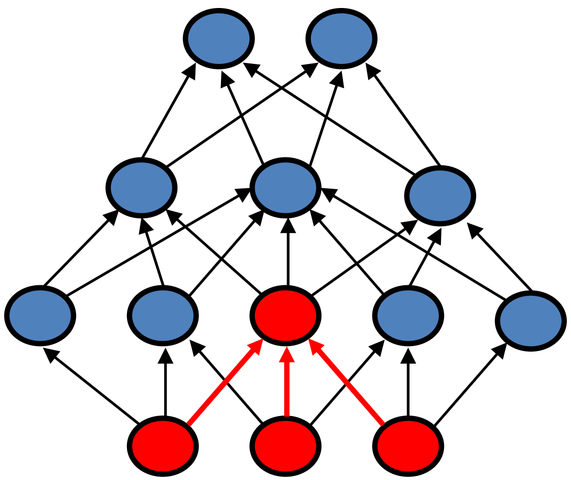
Neural network

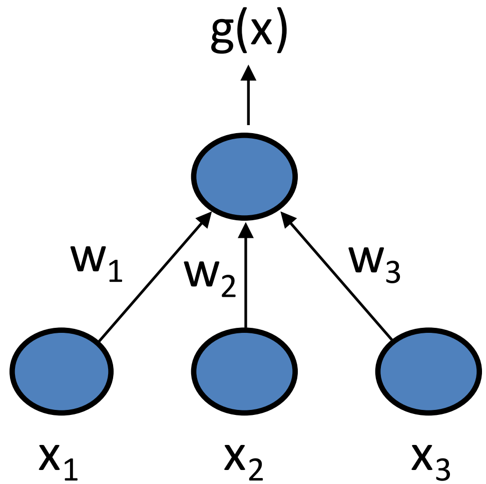
Back propagation

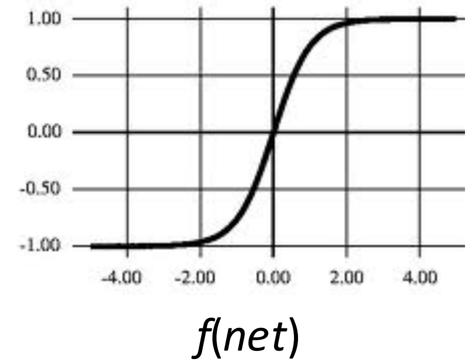
Nature

1940s

1986







$$g(\mathbf{x}) = f\left(\sum_{i=1}^{d} x_i w_i + w_0\right) = f(\mathbf{w}^t\mathbf{x})$$

Lectures

Week	Topics	Requirements
1 (Jan 10 & 12)	Introduction	
2 (Jan 17 & 19)	Machine learning basics	
3 (Jan 24 & 26)	Multilayer neural networks	Homework 1
	Chinese New Year	
4 (Feb 7 & 9)	Convolutional neural netowrks	Homework 2
5 (Feb 14 & 16)	**Optimization for training deep neural networks**	
6 (Feb 21 & 23)	Network structures	Quiz 1 (Feb 21)
7 (Feb 28 & Mar 2)	Recurrent neural network (RNN) and LSTM	
8 (Mar 7 & 9)	Deep belief net and auto-encoder	Homework 3
9 (Mar 14 & 16)	Reinforcement learning & deep learning	Project proposal
10 (Mar 21 & 23)	Attention models	
11 (Mar 28 & 30)	Generative adversarial networks (GAN)	
12 (Apr 4 & 6)	Structured deep learning	Quiz 2 (Apr 4)
13 (Apr 11 & 18)	Course sum-up	
	Project presentation (to be decided)	

Neural network

↓

1940s

Back propagation

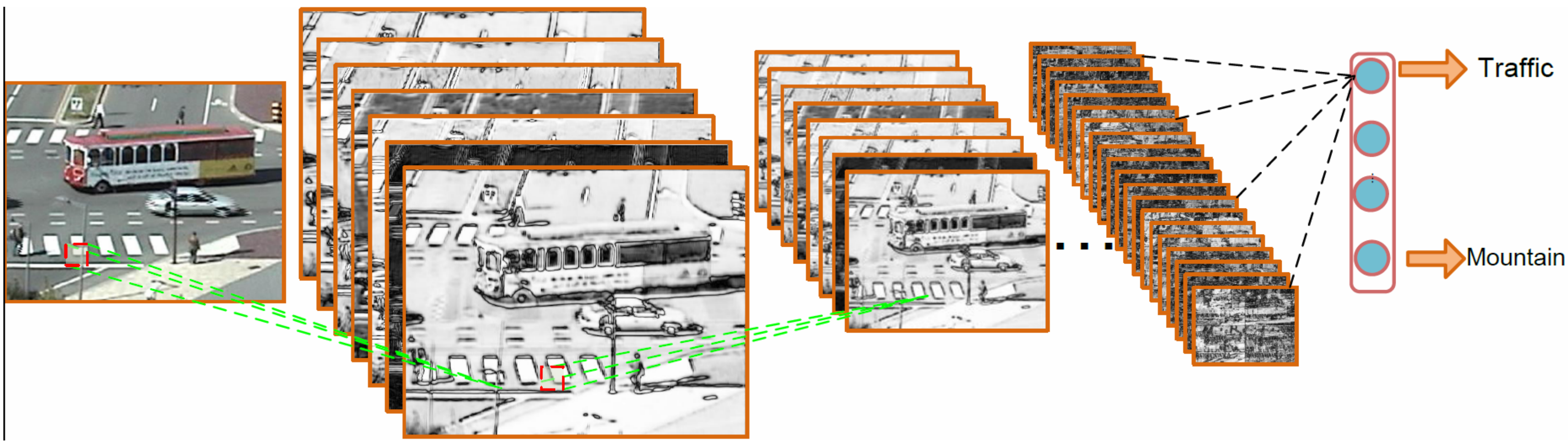
↓

1986

Convolutional neural network

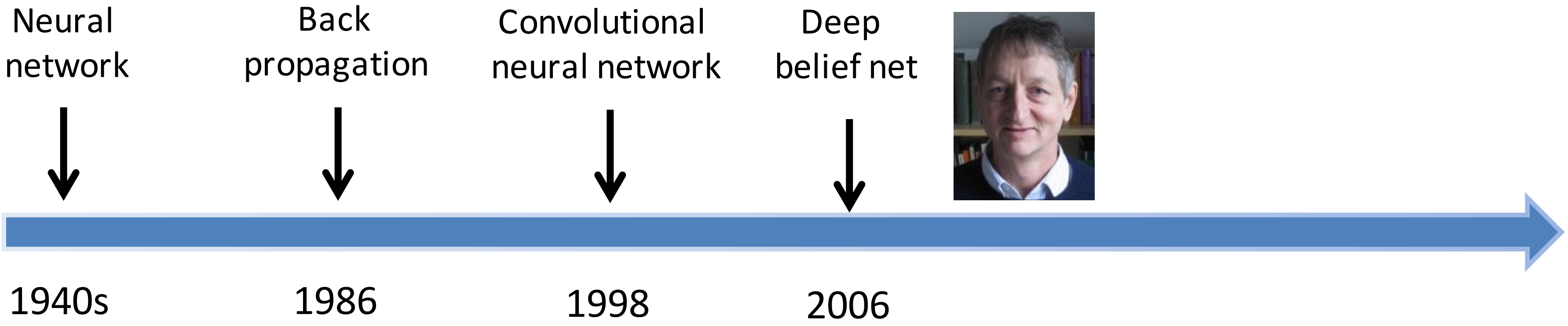
↓

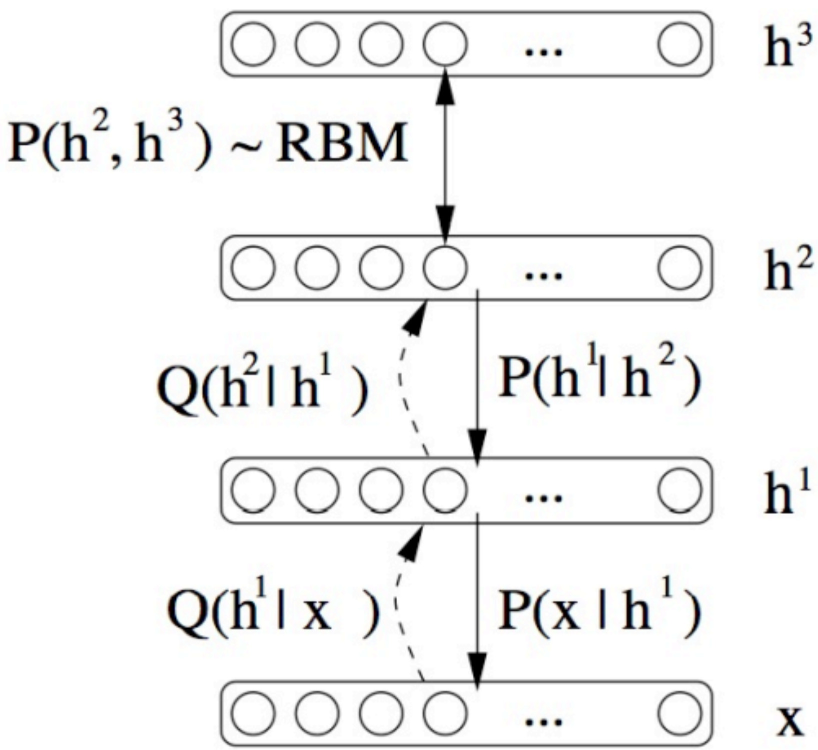
1998



Lectures

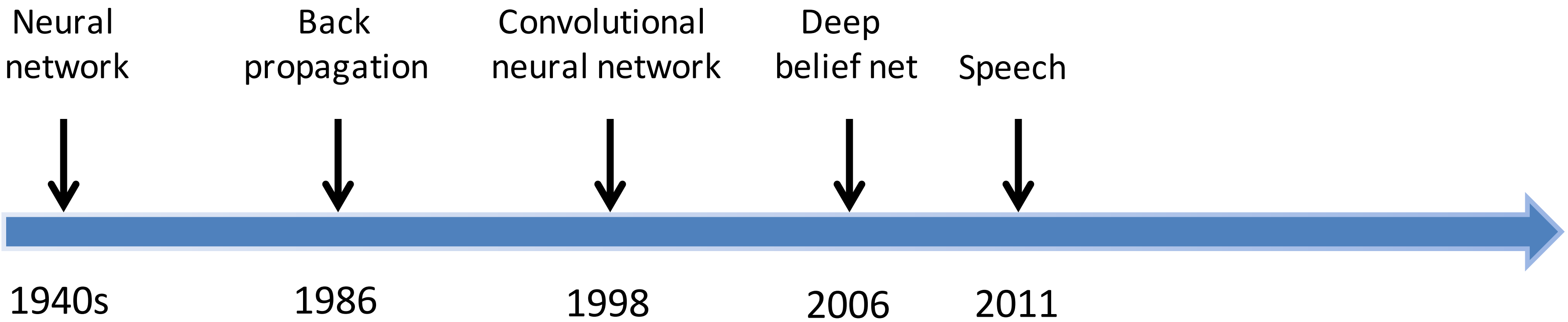
Week	Topics	Requirements
1 (Jan 10 & 12)	Introduction	
2 (Jan 17 & 19)	Machine learning basics	
3 (Jan 24 & 26)	Multilayer neural networks	Homework 1
	Chinese New Year	
4 (Feb 7 & 9)	**Convolutional neural netowrks**	**Homework 2**
5 (Feb 14 & 16)	Optimization for training deep neural networks	
6 (Feb 21 & 23)	Network structures	Quiz 1 (Feb 21)
7 (Feb 28 & Mar 2)	Recurrent neural network (RNN) and LSTM	
8 (Mar 7 & 9)	Deep belief net and auto-encoder	Homework 3
9 (Mar 14 & 16)	Reinforcement learning & deep learning	Project proposal
10 (Mar 21 & 23)	Attention models	
11 (Mar 28 & 30)	Generative adversarial networks (GAN)	
12 (Apr 4 & 6)	Structured deep learning	Quiz 2 (Apr 4)
13 (Apr 11 & 18)	Course sum-up	
	Project presentation (to be decided)	


Neural network
Back propagation
Convolutional neural network
Deep belief net
1940s
1986
1998
2006


h3
h2
h1
x
P(h2, h3) ~ RBM
Q(h2 | h1)
P(h1 | h2)
Q(h1 | x)
P(x | h1)

Lectures

Week	Topics	Requirements
1 (Jan 10 & 12)	Introduction	
2 (Jan 17 & 19)	Machine learning basics	
3 (Jan 24 & 26)	Multilayer neural networks	Homework 1
Chinese New Year		
4 (Feb 7 & 9)	Convolutional neural netowrks	Homework 2
5 (Feb 14 & 16)	Optimization for training deep neural networks	
6 (Feb 21 & 23)	Network structures	Quiz 1 (Feb 21)
7 (Feb 28 & Mar 2)	Recurrent neural network (RNN) and LSTM	
8 (Mar 7 & 9)	**Deep belief net and auto-encoder**	**Homework 3**
9 (Mar 14 & 16)	Reinforcement learning & deep learning	Project proposal
10 (Mar 21 & 23)	Attention models	
11 (Mar 28 & 30)	Generative adversarial networks (GAN)	
12 (Apr 4 & 6)	Structured deep learning	Quiz 2 (Apr 4)
13 (Apr 11 & 18)	Course sum-up	
Project presentation (to be decided)		

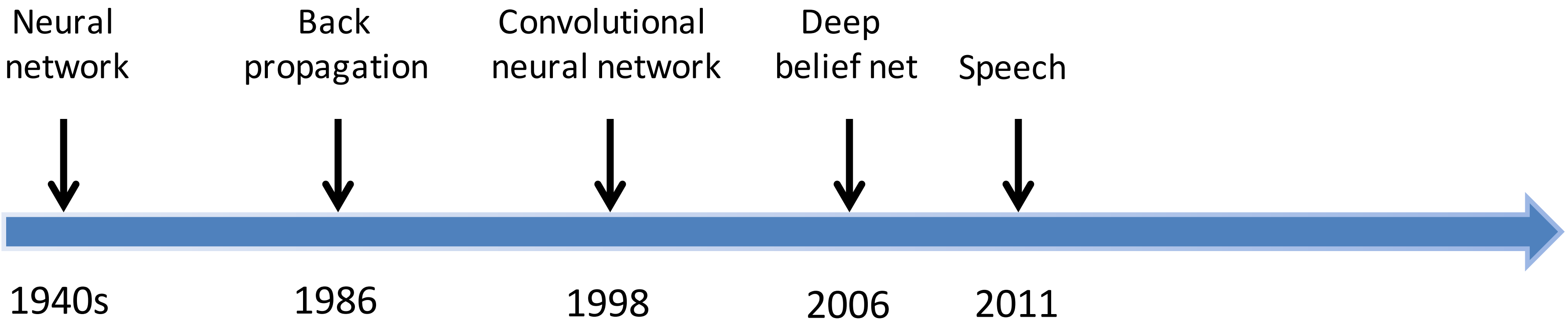


deep learning results

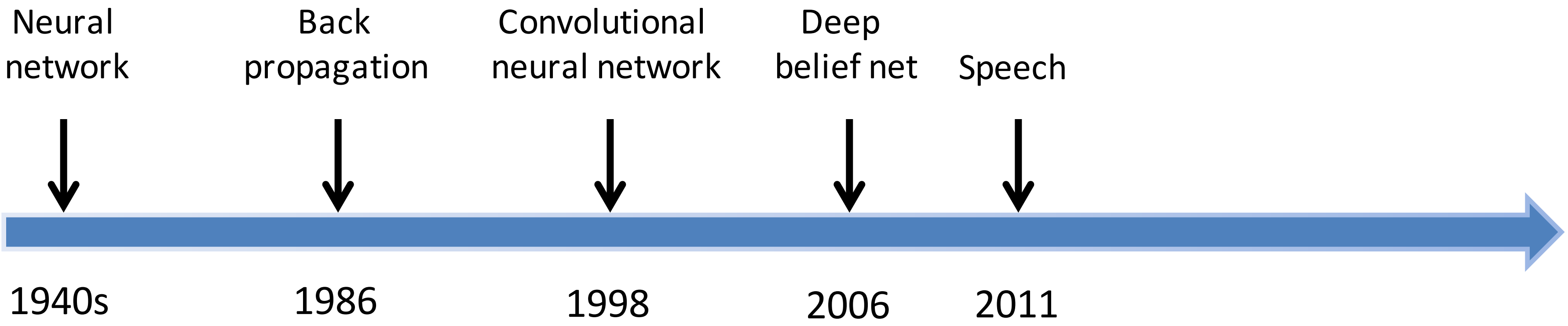
task	hours of training data	DNN-HMM	GMM-HMM with same data
Switchboard (test set 1)	309	18.5	27.4
Switchboard (test set 2)	309	16.1	23.6
English Broadcast News	50	17.5	18.8
Bing Voice Search (Sentence error rates)	24	30.4	36.2
Google Voice Input	5,870	12.3	
Youtube	1,400	47.6	52.3

Deep Networks Advance State of Art in Speech

Deep Learning leads to breakthrough in speech recognition at MSR.



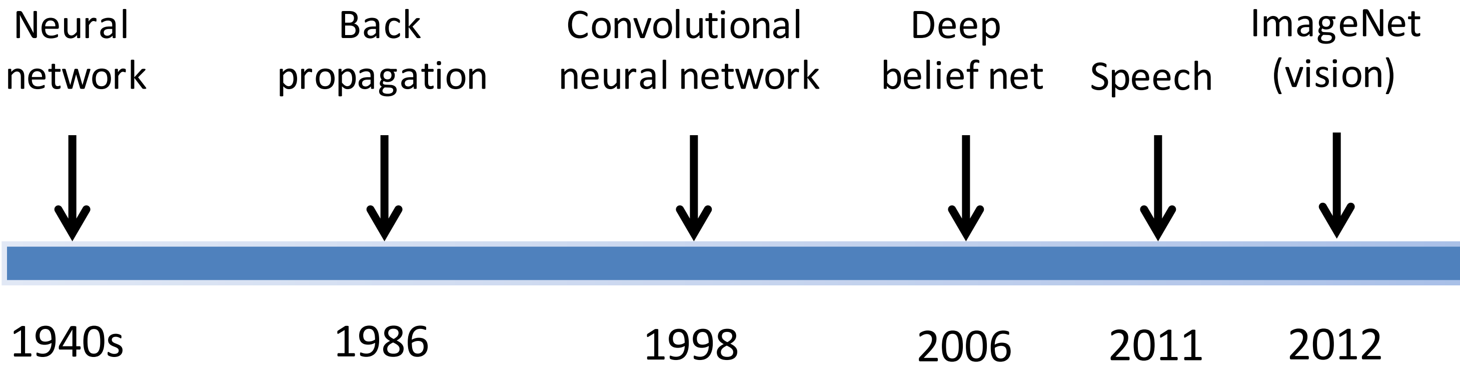
Not well accepted by the vision community ☹



LeCun's open letter in CVPR 2012

So, I'm giving up on submitting to computer vision conferences altogether. CV reviewers are just too likely to be clueless or hostile towards our brand of methods. Submitting our papers is just a waste of everyone's time (and incredibly demoralizing to my lab members)

I might come back in a few years, if at least two things change:
- Enough people in CV become interested in feature learning that the probability of getting a non-clueless and non-hostile reviewer is more than 50% (hopefully [Computer Vision Researcher]'s tutorial on the topic at CVPR will have some positive effect).
- CV conference proceedings become open access.

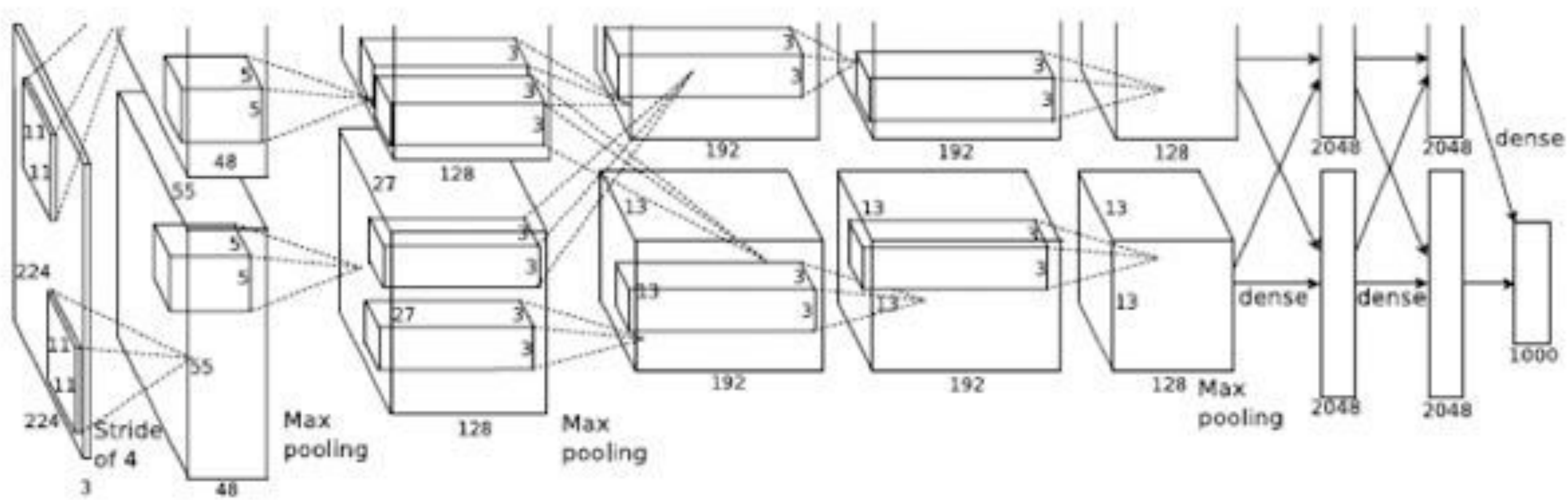


Rank	Name	Error rate	Description
1	**U. Toronto**	0.15315	Deep learning
2	U. Tokyo	0.26172	Hand-crafted features and learning models. Bottleneck.
3	U. Oxford	0.26979	
4	Xerox/INRIA	0.27058	

Object recognition over 1,000,000 images and 1,000 categories (2 GPU)

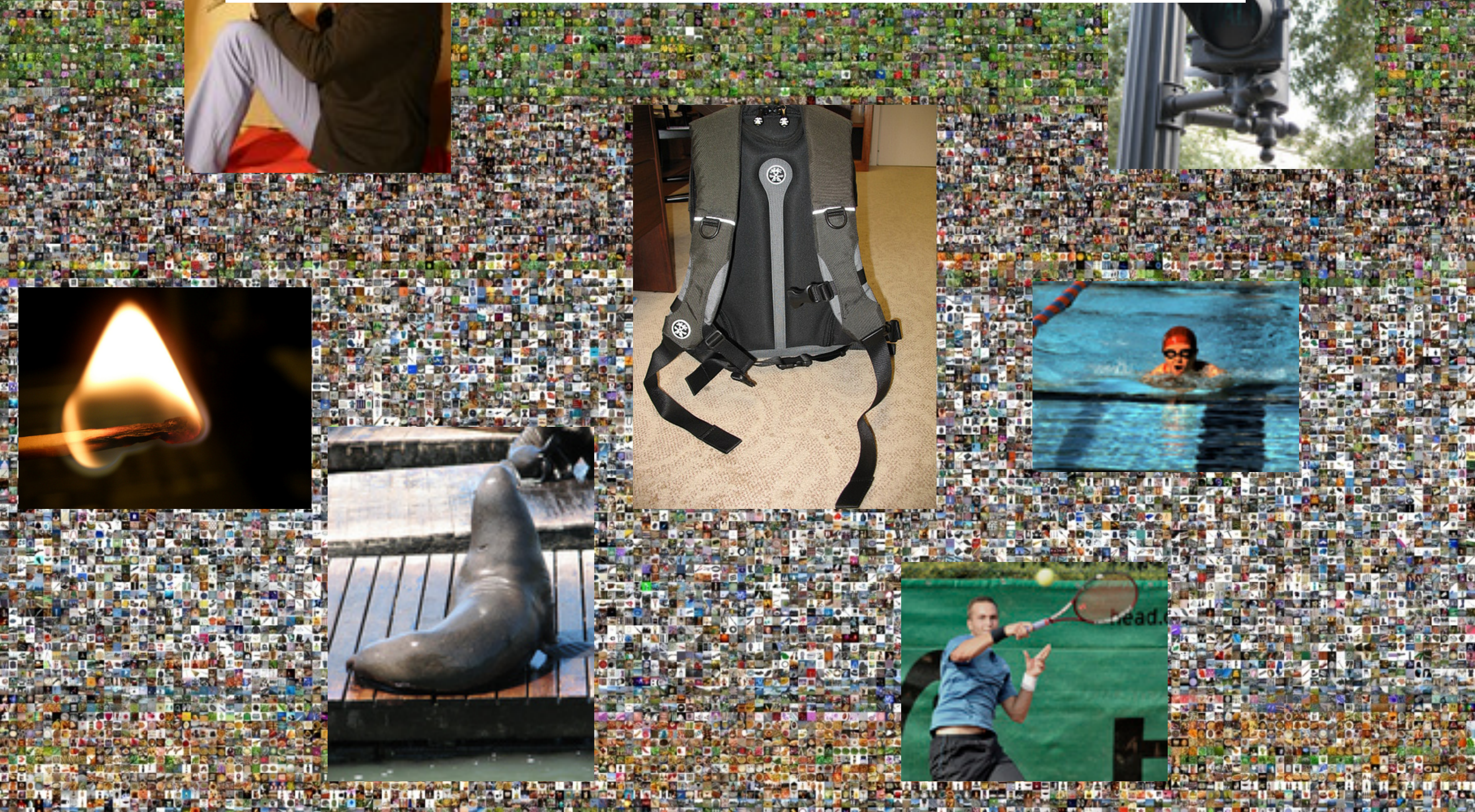
Current best result < 0.03

A. Krizhevsky, L. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," NIPS, 2012.



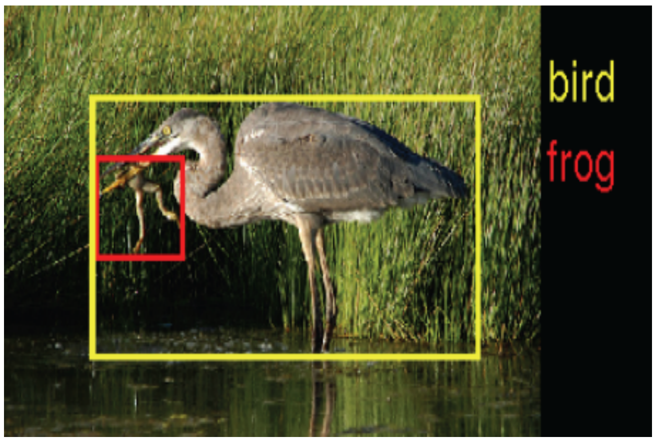
AlexNet implemented on 2 GPUs (each has 2GB memory)

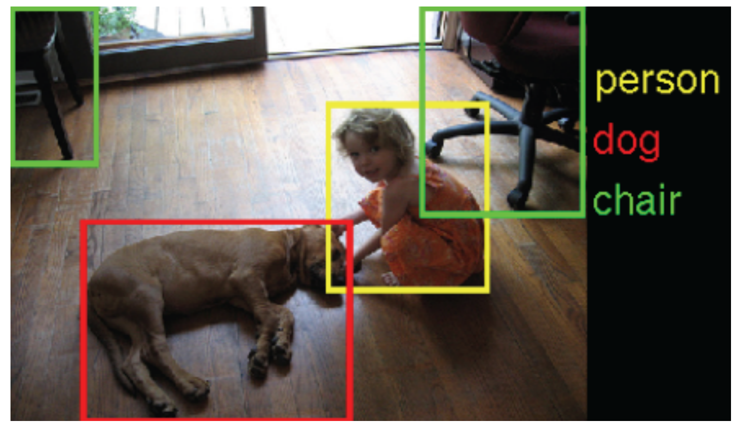
ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

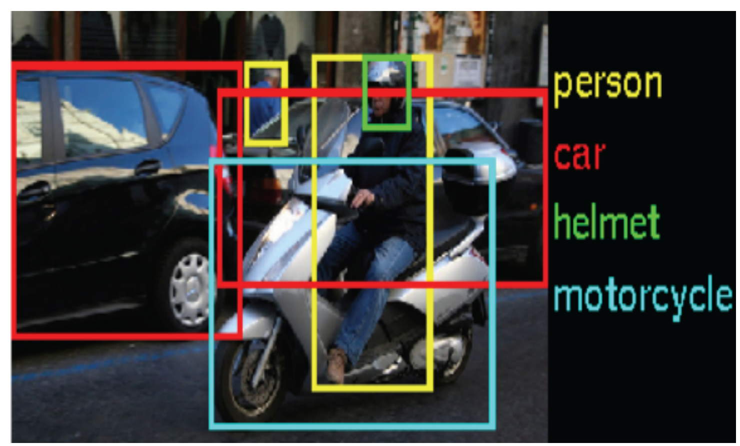


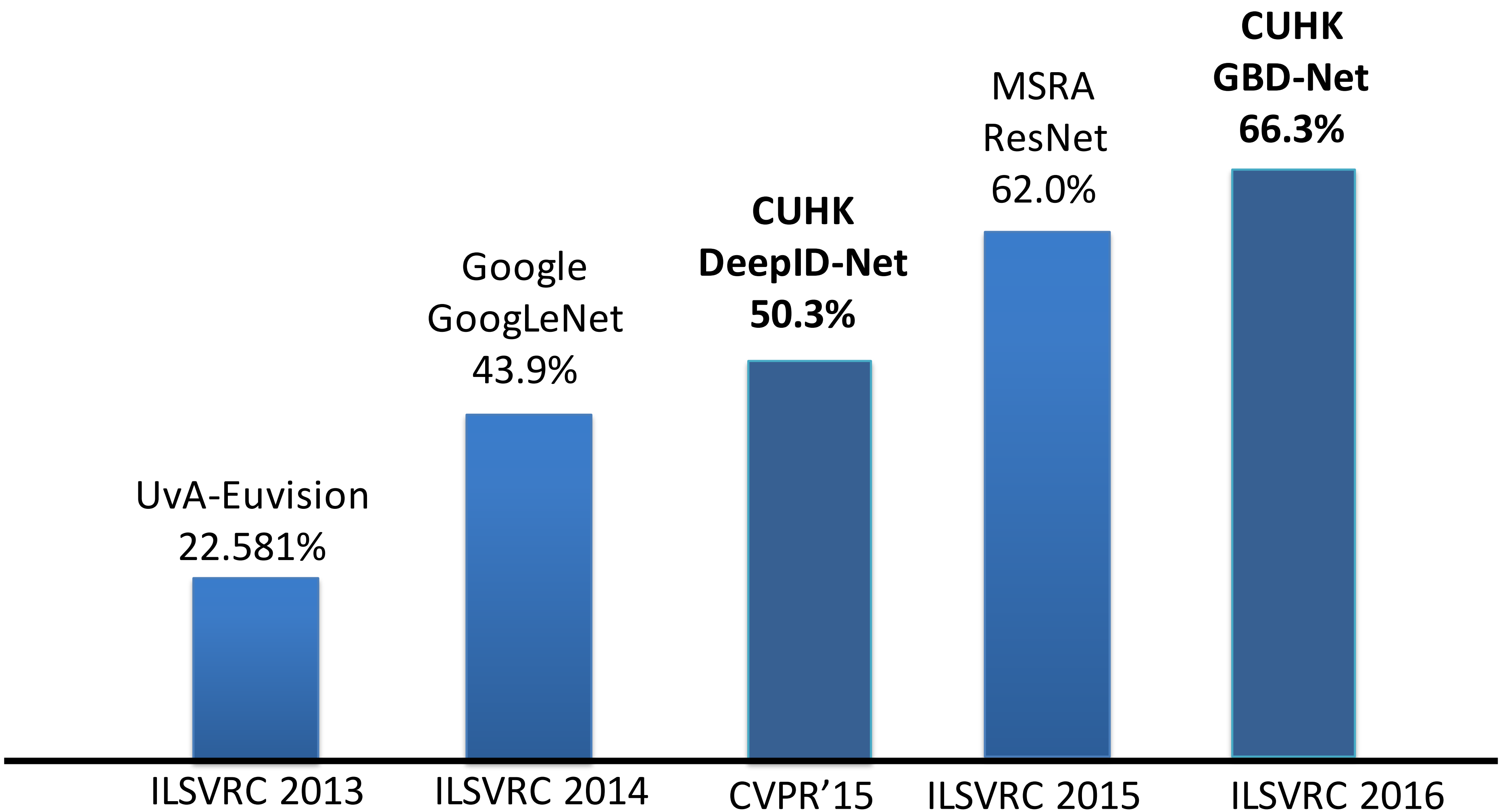
ImageNet Object Detection Task

- 200 object classes
- 60,000 test images

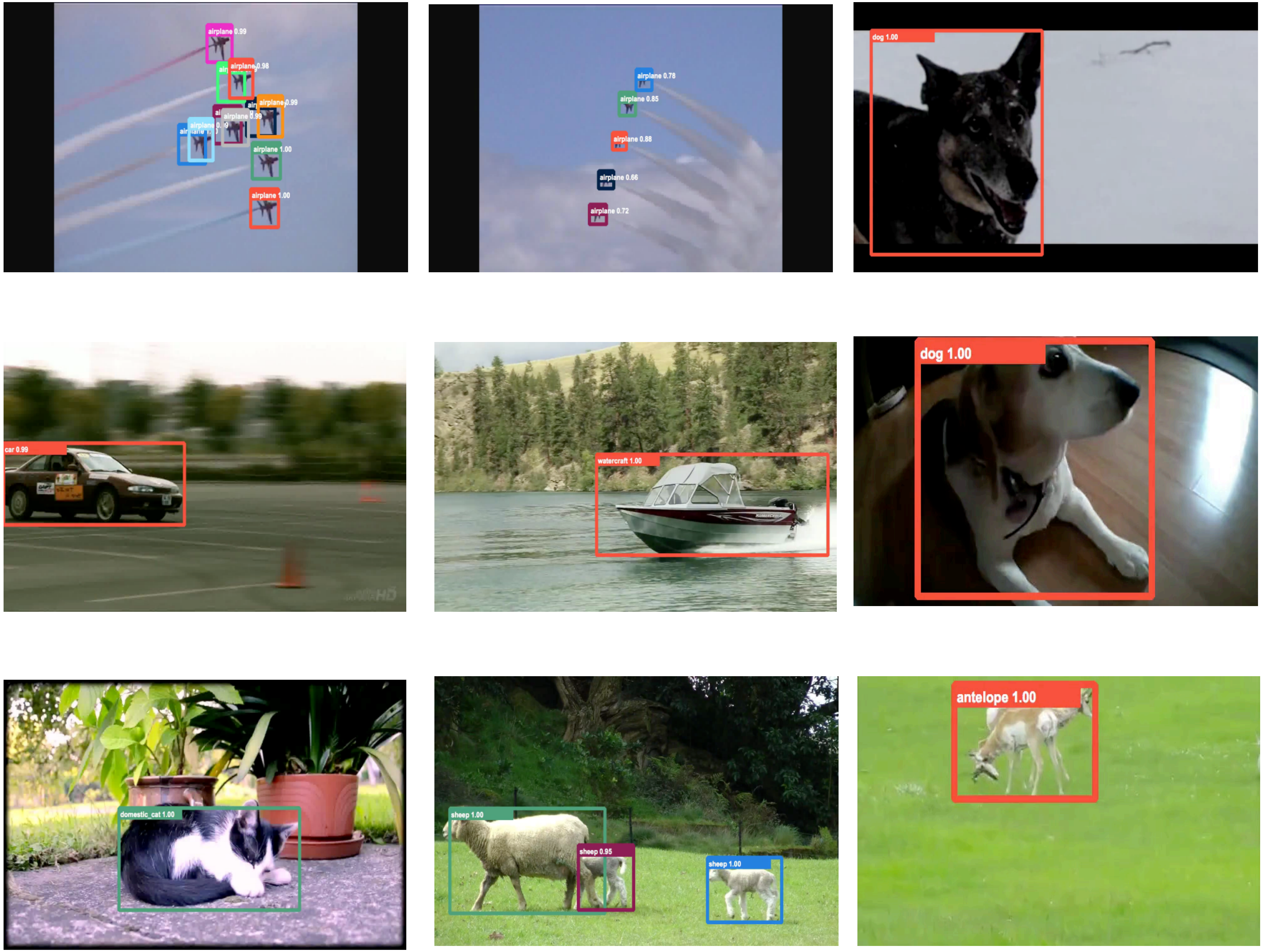




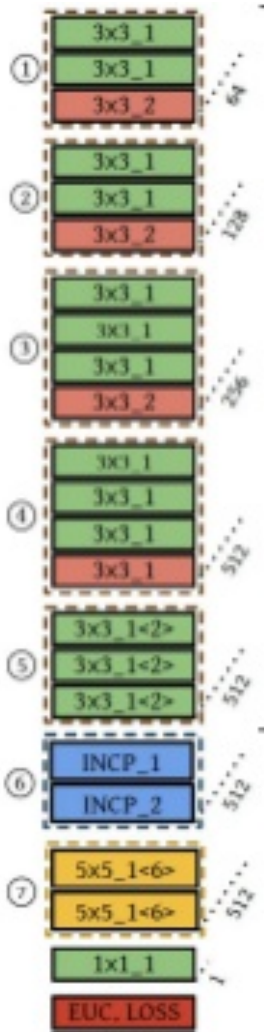




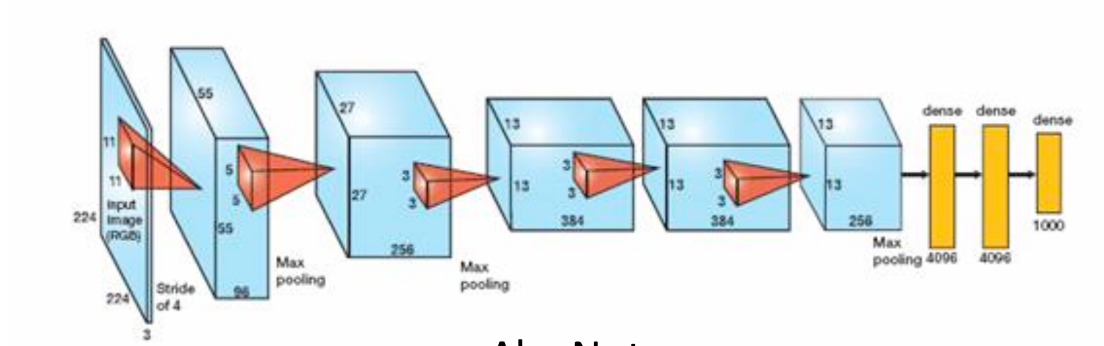
UvA-Euvision
22.581%
Google
GoogLeNet
43.9%
CUHK
DeepID-Net
50.3%
MSRA
ResNet
62.0%
CUHK
GBD-Net
66.3%
ILSVRC 2013
ILSVRC 2014
CVPR'15
ILSVRC 2015
ILSVRC 2016


airplane 0.99
airplane 0.98
airplane 0.99
airplane 1.00
airplane 1.00
airplane 0.78
airplane 0.85
airplane 0.88
airplane 0.66
airplane 0.72
dog 1.00
car 0.99
watercraft 1.00
dog 1.00
domestic_cat 1.00
sheep 1.00
sheep 0.95
sheep 1.00
antelope 1.00

Network Structures

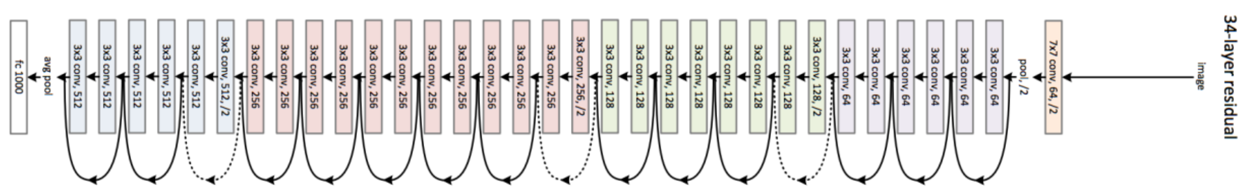


VGG



AlexNet

GoogLeNet



ResNet

Lectures

Week	Topics	Requirements
1 (Jan 10 & 12)	Introduction	
2 (Jan 17 & 19)	Machine learning basics	
3 (Jan 24 & 26)	Multilayer neural networks	Homework 1
	Chinese New Year	
4 (Feb 7 & 9)	Convolutional neural netowrks	Homework 2
5 (Feb 14 & 16)	Optimization for training deep neural networks	
6 (Feb 21 & 23)	**Network structures**	**Quiz 1 (Feb 21)**
7 (Feb 28 & Mar 2)	Recurrent neural network (RNN) and LSTM	
8 (Mar 7 & 9)	Deep belief net and auto-encoder	Homework 3
9 (Mar 14 & 16)	Reinforcement learning & deep learning	Project proposal
10 (Mar 21 & 23)	Attention models	
11 (Mar 28 & 30)	Generative adversarial networks (GAN)	
12 (Apr 4 & 6)	Structured deep learning	Quiz 2 (Apr 4)
13 (Apr 11 & 18)	Course sum-up	
	Project presentation (to be decided)	

Deep Learning Frameworks

Caffe

Theano

Torch

TensorFlow

mxnet

Tutorials

Times	Topic
1	Python/Numpy tutorial/AWS tutorial
2	Understand backpropagation
3	**Torch tutorial**
4	**Caffe/Tensorflow/Theano**
5	Roadmaps of deep learning models
6	Hands on experiment with debugging models
7	GPU parallel programming
8	Final project proposal discussion
9	Assignment and quiz review
10	Fancy stuff: deep learning on spark, future directions

Hands-on assignments are provided in tutorials. Bring your laptop

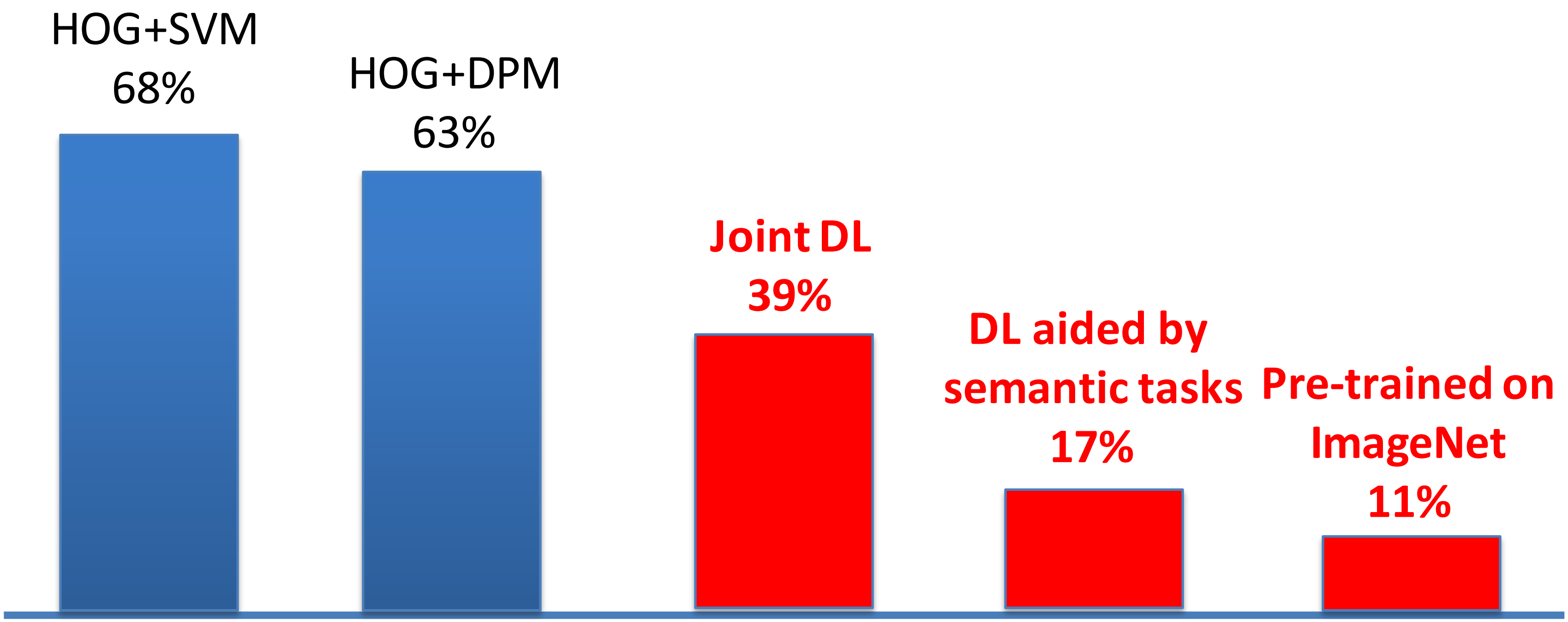
Pedestrian Detection

LatSVM-V2 | LatSVM-V2+Our

False positive detected by LatSVM-V2, but not ours

True positives detected by ours but not LatSVM-V2

Pedestrian detection on Caltech (average miss detection rates)



W. Ouyang and X. Wang, "Joint Deep Learning for Pedestrian Detection," ICCV 2013.

Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian Detection aided by Deep Learning Semantic Tasks," CVPR 2015.

Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep Learning Strong Parts for Pedestrian Detection," ICCV 2015.

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

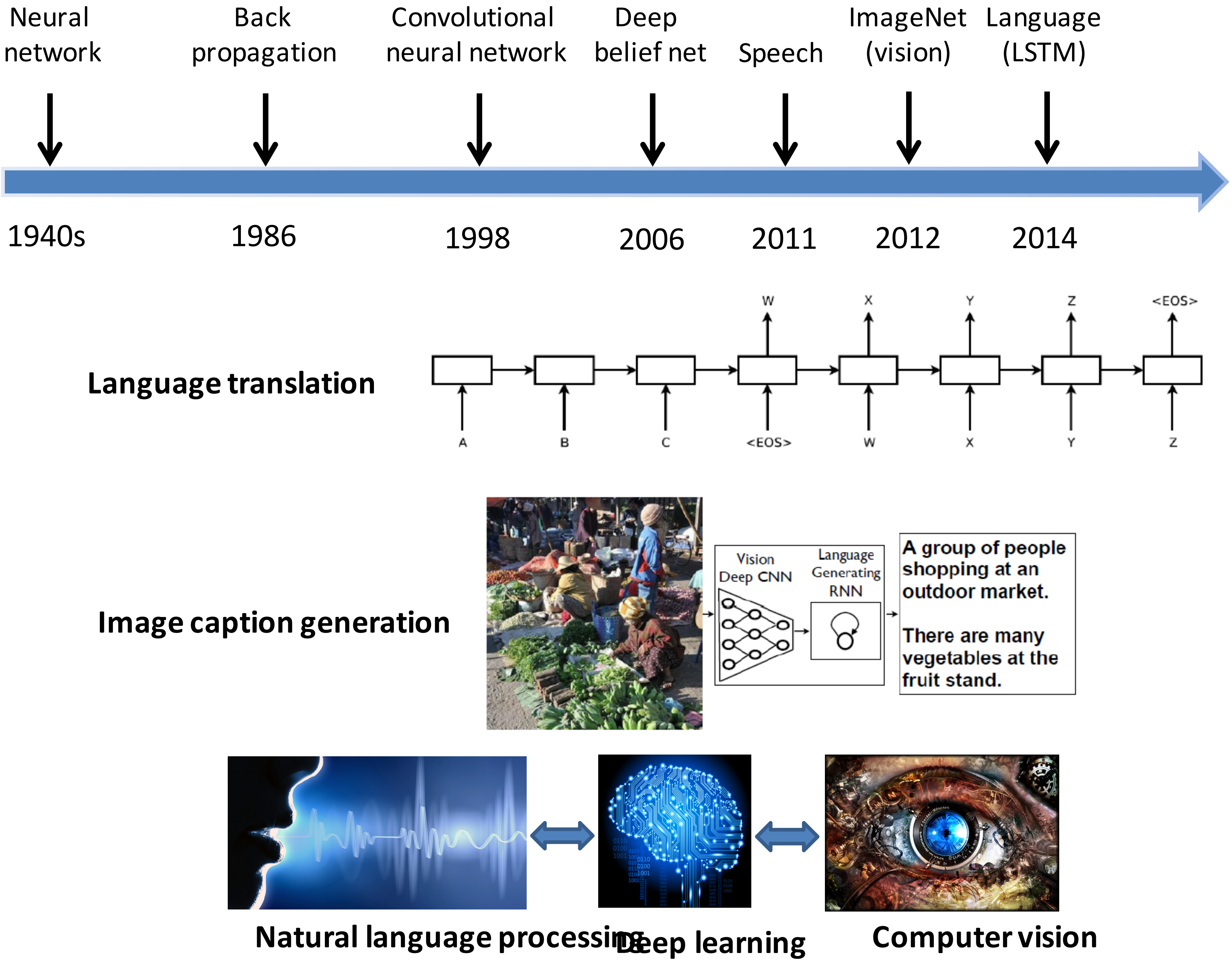
Doubling the efficiency of a solar cell would completely change the economics of renewable energy. Nanotechnology just might make it possible.

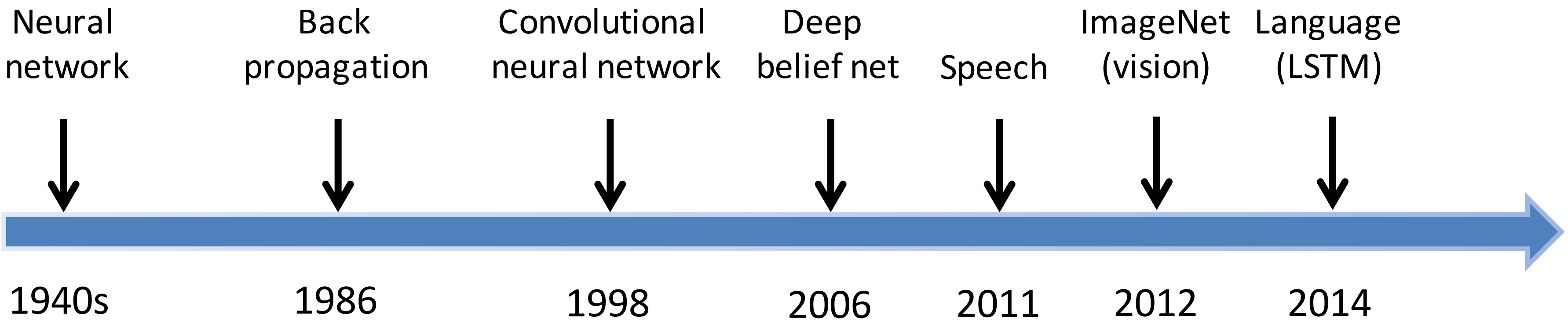
Big Data from Cheap Phones

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.

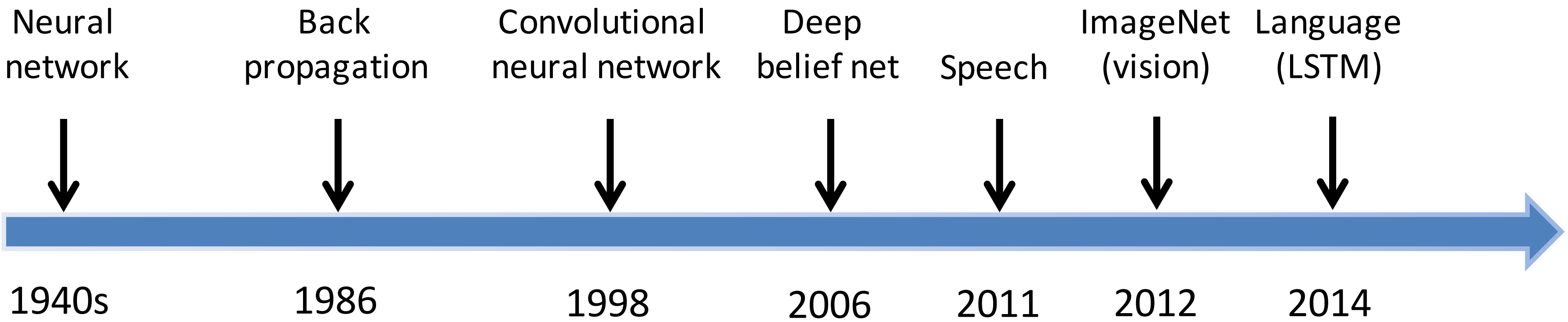

Neural network
Back propagation
Convolutional neural network
Deep belief net
Speech
ImageNet (vision)
Language (LSTM)
1940s
1986
1998
2006
2011
2012
2014
Language translation
W
X
Y
Z
<EOS>
A
B
C
<EOS>
W
X
Y
Z
Image caption generation
Vision Deep CNN
Language Generating RNN
A group of people shopping at an outdoor market.
There are many vegetables at the fruit stand.
Natural language processing
Deep learning
Computer vision



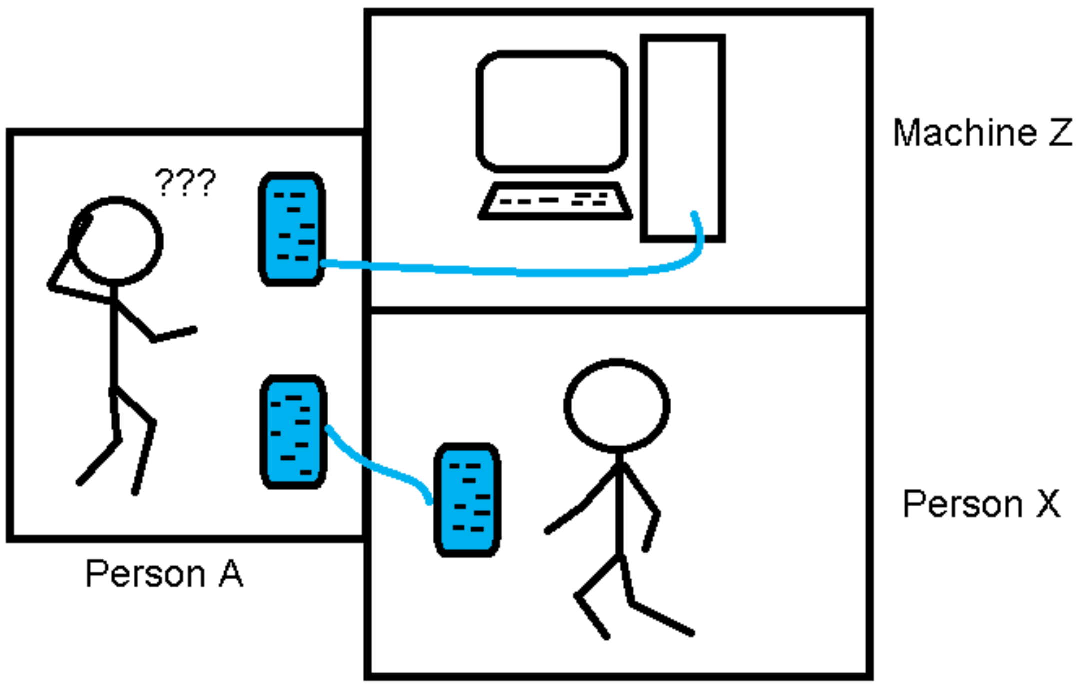
ChatBot

Siris

Xiao Bing



Turing test



Strong AI

Weak AI

Lectures

Week	Topics	Requirements
1 (Jan 10 & 12)	Introduction	
2 (Jan 17 & 19)	Machine learning basics	
3 (Jan 24 & 26)	Multilayer neural networks	Homework 1
Chinese New Year		
4 (Feb 7 & 9)	Convolutional neural netowrks	Homework 2
5 (Feb 14 & 16)	Optimization for training deep neural networks	
6 (Feb 21 & 23)	Network structures	Quiz 1 (Feb 21)
7 (Feb 28 & Mar 2)	**Recurrent neural network (RNN) and LSTM**	
8 (Mar 7 & 9)	Deep belief net and auto-encoder	Homework 3
9 (Mar 14 & 16)	Reinforcement learning & deep learning	Project proposal
10 (Mar 21 & 23)	Attention models	
11 (Mar 28 & 30)	Generative adversarial networks (GAN)	
12 (Apr 4 & 6)	Structured deep learning	Quiz 2 (Apr 4)
13 (Apr 11 & 18)	Course sum-up	
Project presentation (to be decided)		

MIT Technology Review

BUSINESS NEWS

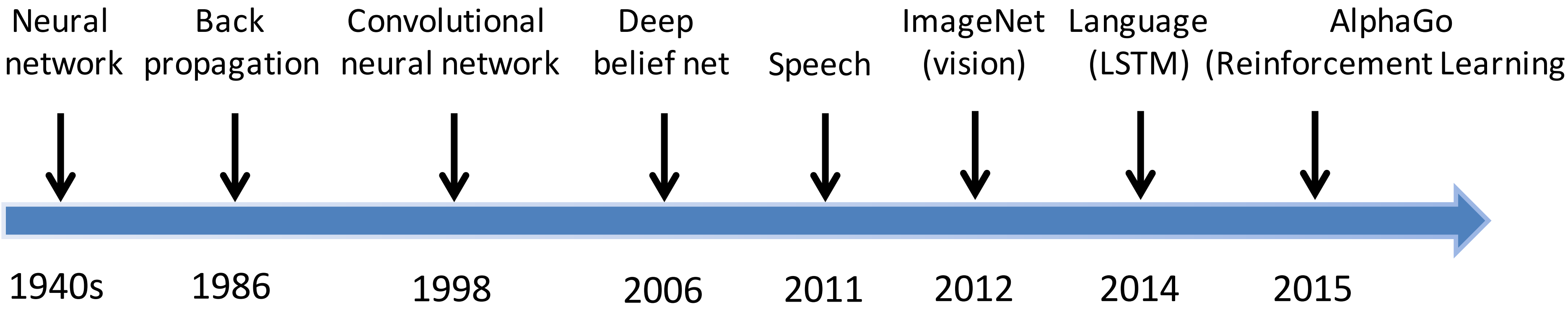
Is Google Cornering the Market on Deep Learning?

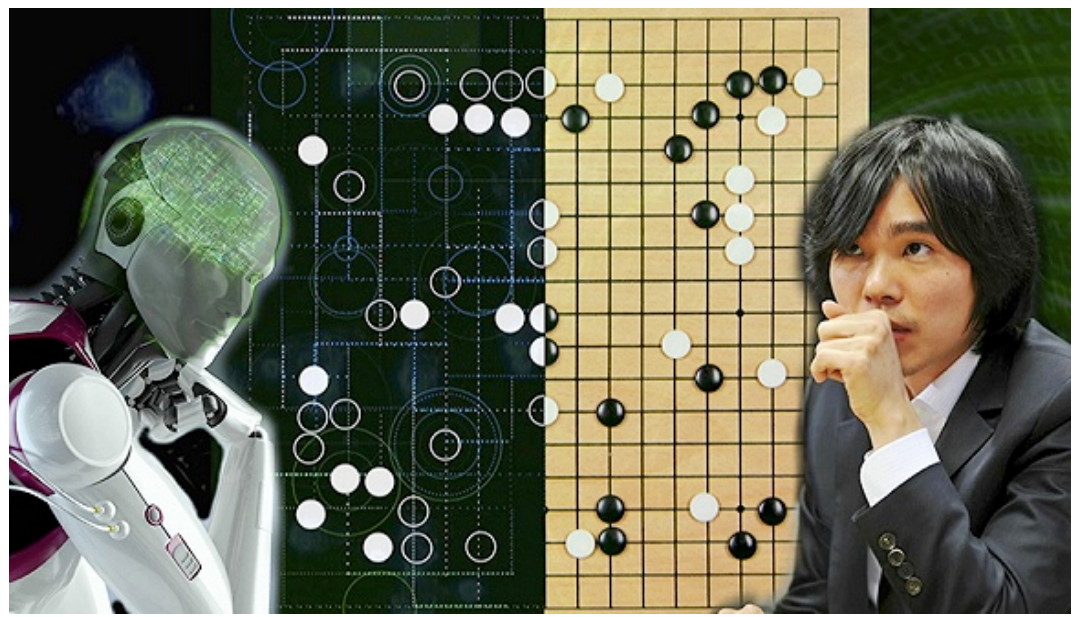
A cutting-edge corner of science is being wooed by Silicon Valley, to the dismay of some academics.

By Antonio Regalado on January 29, 2014

How much are a dozen deep-learning researchers worth? Apparently, more than $400 million.

Yoshua Bengio, an AI researcher at the University of Montreal, **estimates that there are only about 50 experts worldwide in deep learning, many of whom are still graduate students**. He estimated that DeepMind employed about a dozen of them on its staff of about 50. “I think this is the main reason that Google bought DeepMind. It has one of the largest concentrations of deep learning experts,” Bengio says.

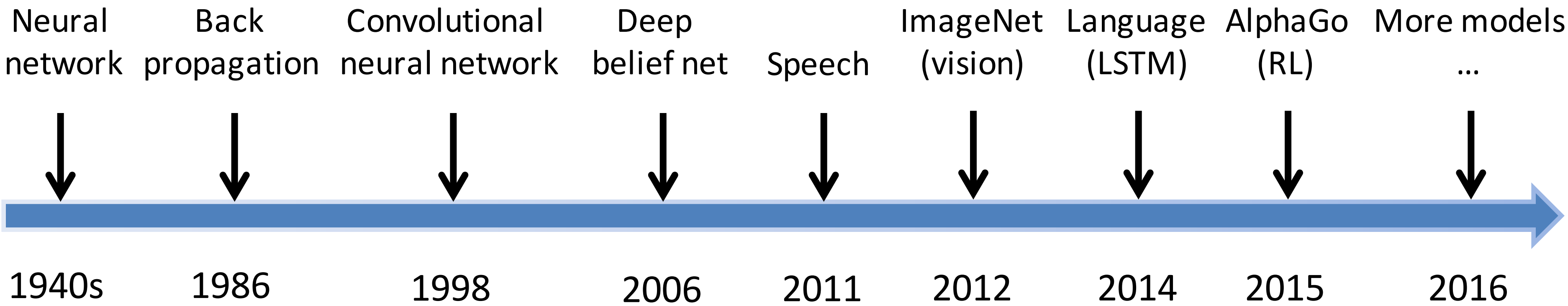




1920 CPU and 280 GPU

Lectures

Week	Topics	Requirements
1 (Jan 10 & 12)	Introduction	
2 (Jan 17 & 19)	Machine learning basics	
3 (Jan 24 & 26)	Multilayer neural networks	Homework 1
	Chinese New Year	
4 (Feb 7 & 9)	Convolutional neural netowrks	Homework 2
5 (Feb 14 & 16)	Optimization for training deep neural networks	
6 (Feb 21 & 23)	Network structures	Quiz 1 (Feb 21)
7 (Feb 28 & Mar 2)	Recurrent neural network (RNN) and LSTM	
8 (Mar 7 & 9)	Deep belief net and auto-encoder	Homework 3
9 (Mar 14 & 16)	**Reinforcement learning & deep learning**	**Project proposal**
10 (Mar 21 & 23)	Attention models	
11 (Mar 28 & 30)	Generative adversarial networks (GAN)	
12 (Apr 4 & 6)	Structured deep learning	Quiz 2 (Apr 4)
13 (Apr 11 & 18)	Course sum-up	
	Project presentation (to be decided)	



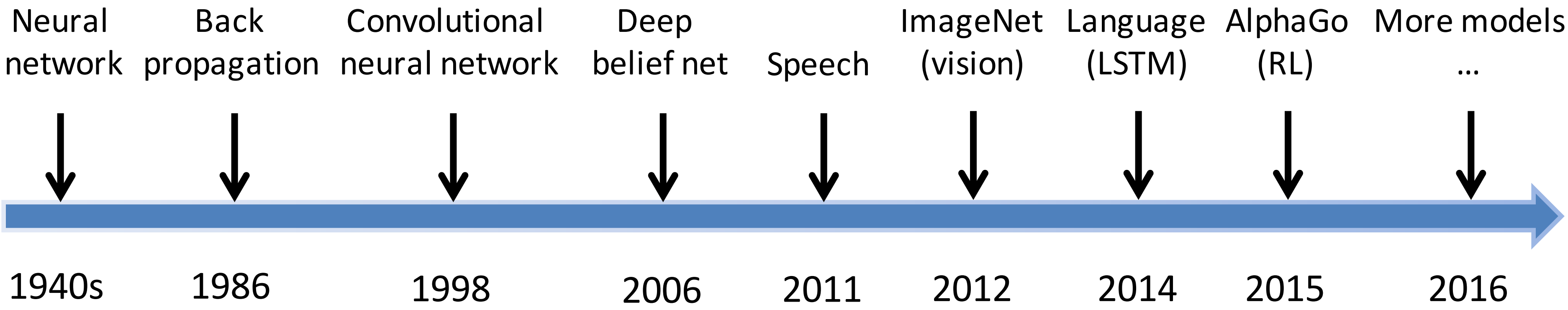
Q: what is the color of the bird? **A**: **white**

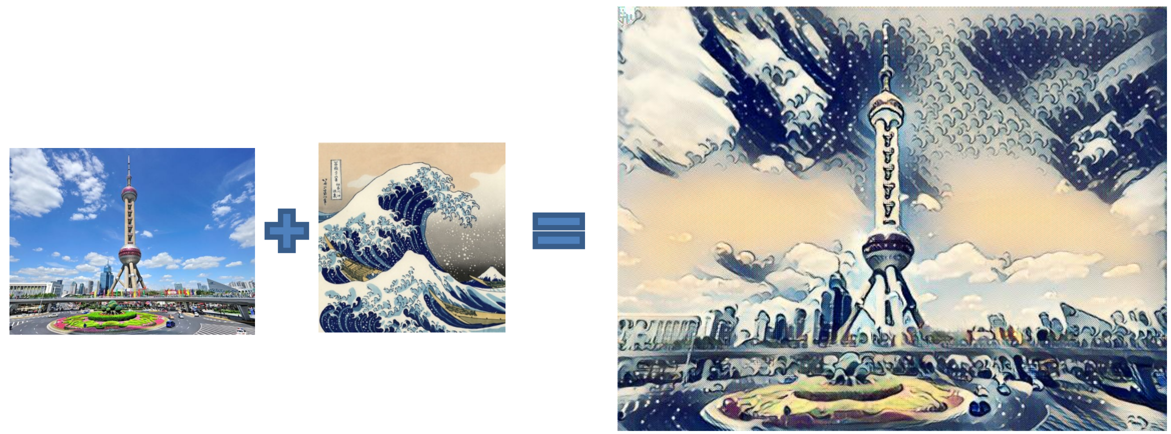
what is the color of the bird ?

Attention models

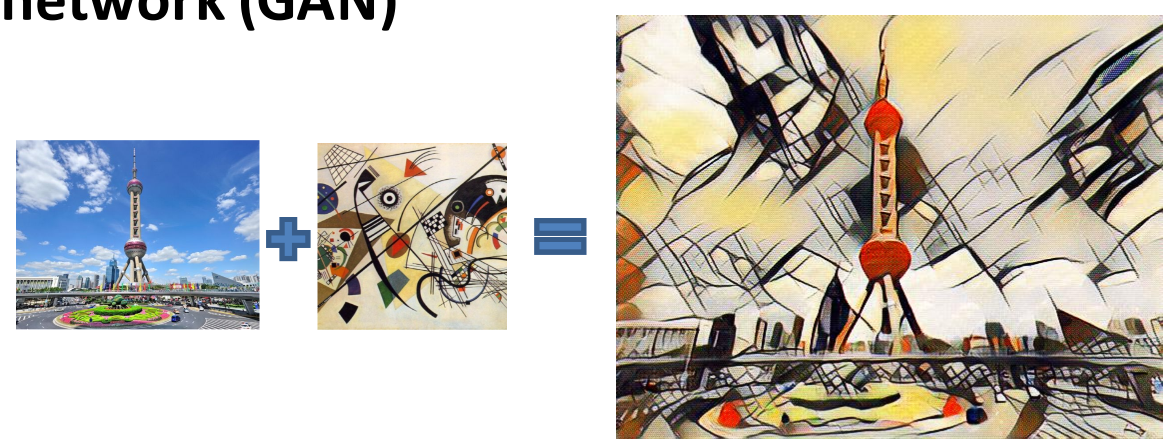
Lectures

Week	Topics	Requirements
1 (Jan 10 & 12)	Introduction	
2 (Jan 17 & 19)	Machine learning basics	
3 (Jan 24 & 26)	Multilayer neural networks	Homework 1
Chinese New Year		
4 (Feb 7 & 9)	Convolutional neural netowrks	Homework 2
5 (Feb 14 & 16)	Optimization for training deep neural networks	
6 (Feb 21 & 23)	Network structures	Quiz 1 (Feb 21)
7 (Feb 28 & Mar 2)	Recurrent neural network (RNN) and LSTM	
8 (Mar 7 & 9)	Deep belief net and auto-encoder	Homework 3
9 (Mar 14 & 16)	Reinforcement learning & deep learning	Project proposal
10 (Mar 21 & 23)	**Attention models**	
11 (Mar 28 & 30)	Generative adversarial networks (GAN)	
12 (Apr 4 & 6)	Structured deep learning	Quiz 2 (Apr 4)
13 (Apr 11 & 18)	Course sum-up	
Project presentation (to be decided)		



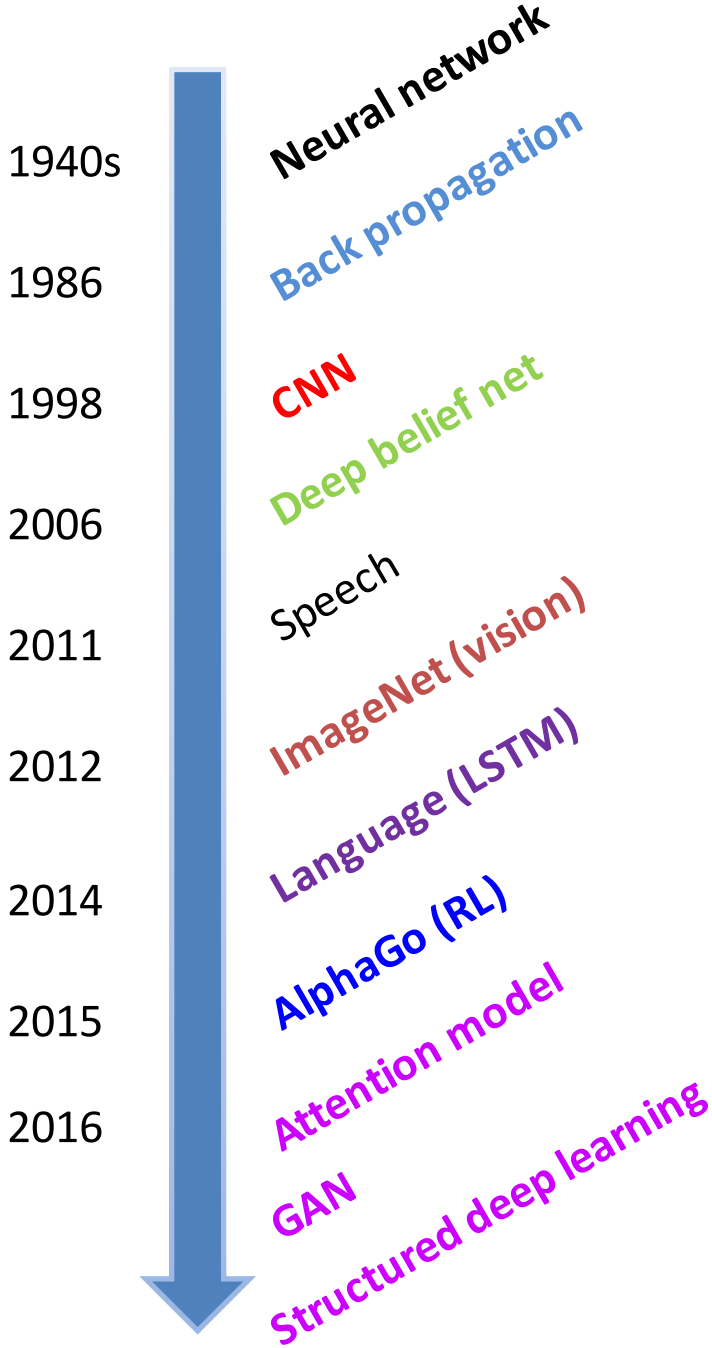


Generative adversarial network (GAN)



Lectures

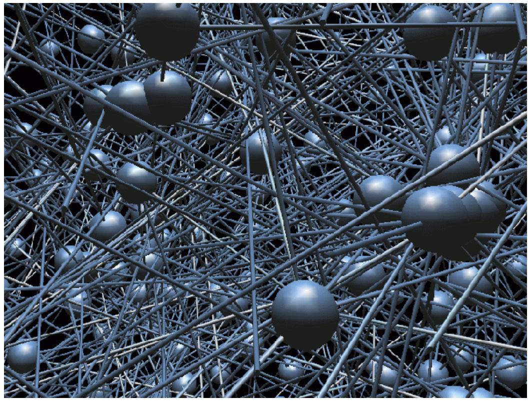
Week	Topics	Requirements
1 (Jan 10 & 12)	Introduction	
2 (Jan 17 & 19)	Machine learning basics	
3 (Jan 24 & 26)	Multilayer neural networks	Homework 1
Chinese New Year		
4 (Feb 7 & 9)	Convolutional neural netowrks	Homework 2
5 (Feb 14 & 16)	Optimization for training deep neural networks	
6 (Feb 21 & 23)	Network structures	Quiz 1 (Feb 21)
7 (Feb 28 & Mar 2)	Recurrent neural network (RNN) and LSTM	
8 (Mar 7 & 9)	Deep belief net and auto-encoder	Homework 3
9 (Mar 14 & 16)	Reinforcement learning & deep learning	Project proposal
10 (Mar 21 & 23)	Attention models	
11 (Mar 28 & 30)	**Generative adversarial networks (GAN)**	
12 (Apr 4 & 6)	Structured deep learning	Quiz 2 (Apr 4)
13 (Apr 11 & 18)	Course sum-up	
Project presentation (to be decided)		



Topics
Introduction
Machine learning basics
Multilayer neural networks
Convolutional neural netowrks
Optimization for training deep neural networks
Network structures
Recurrent neural network (RNN) and LSTM
Deep belief net and auto-encoder
Reinforcement learning & deep learning
Attention models
Generative adversarial networks (GAN)
Structured deep learning
Course sum-up

Outline

- Historical review of deep learning
- **Understand deep learning**
- Interpret Neural Semantics



Highly complex neural networks with many layers, millions or billions of neurons, and sophisticated architectures

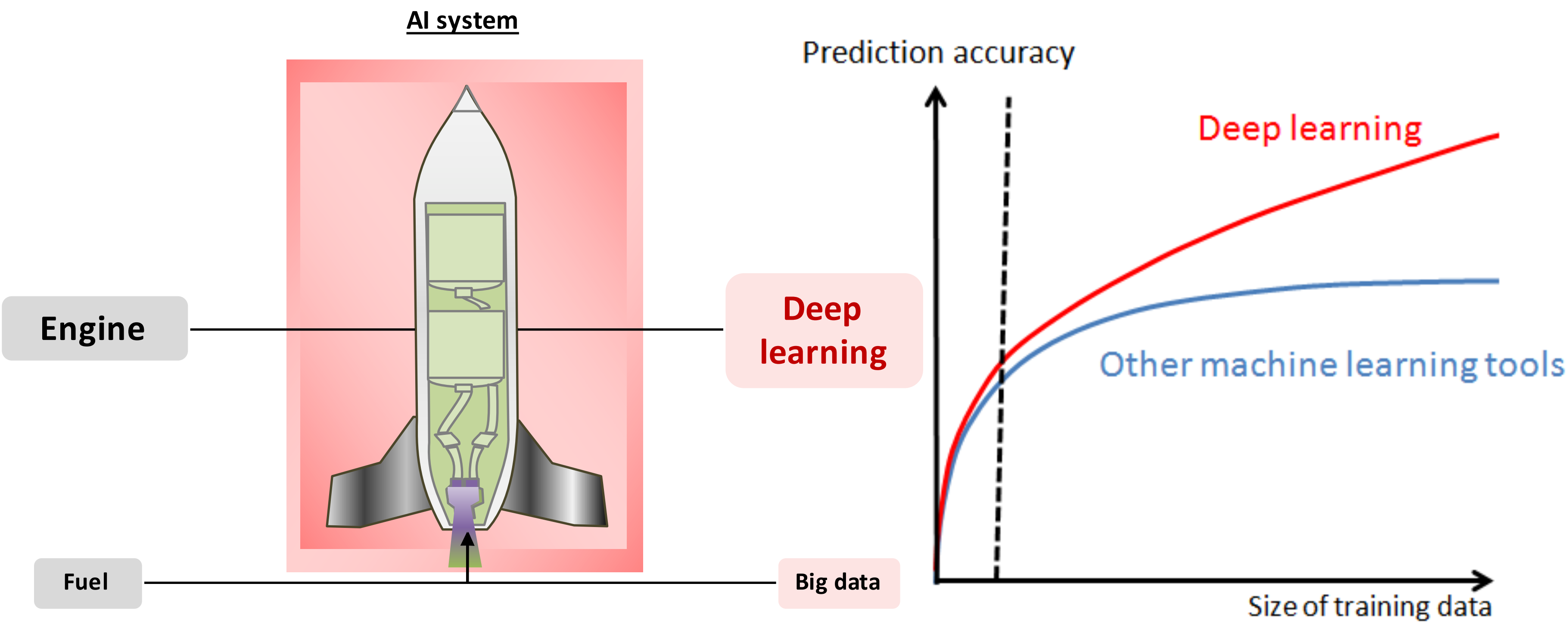
Fit billions of training samples

Trained with GPU clusters with millions of processors

Deep learning

Machine Learning with Big Data

- Machine learning with small data: **overfitting**, reducing model complexity (capacity), adding regularization
- Machine learning with big data: **underfitting**, increasing model complexity, optimization, computation resource



Pattern Recognition = Feature + Classifier

<u>Feature Learning</u> vs Feature Engineering

Deep Learning

Pattern Recognition System

Input

↓

sensing

↓

preprocessing

↓

feature extraction

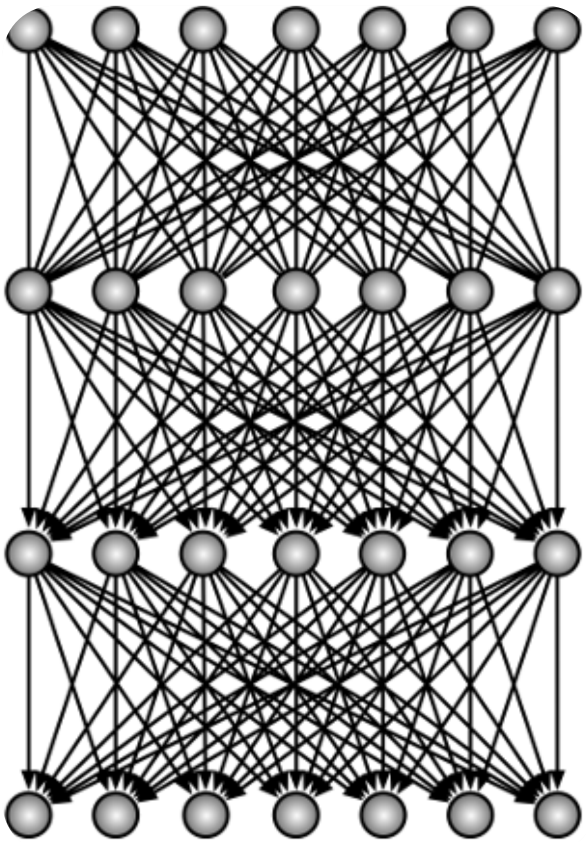
↓

classification

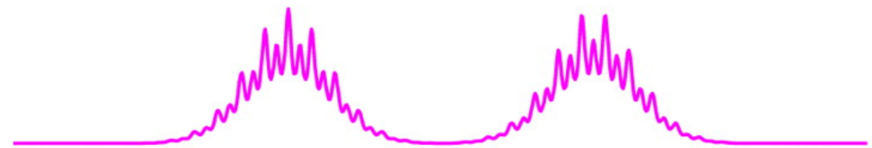
↓

Decision: "salmon" or "sea bass"

Neural Responses are Features



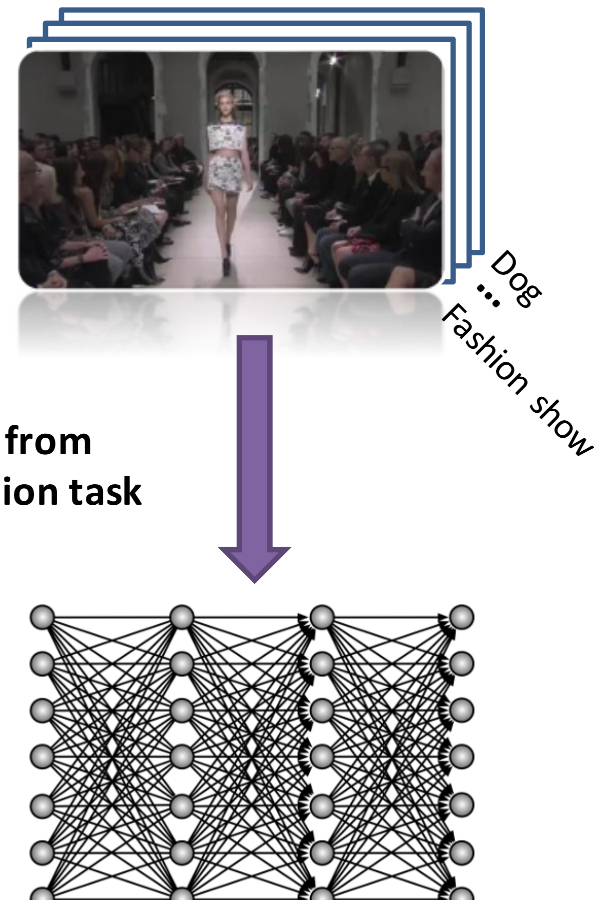
Artificial neural network



Human brain

Way to Learn Features?

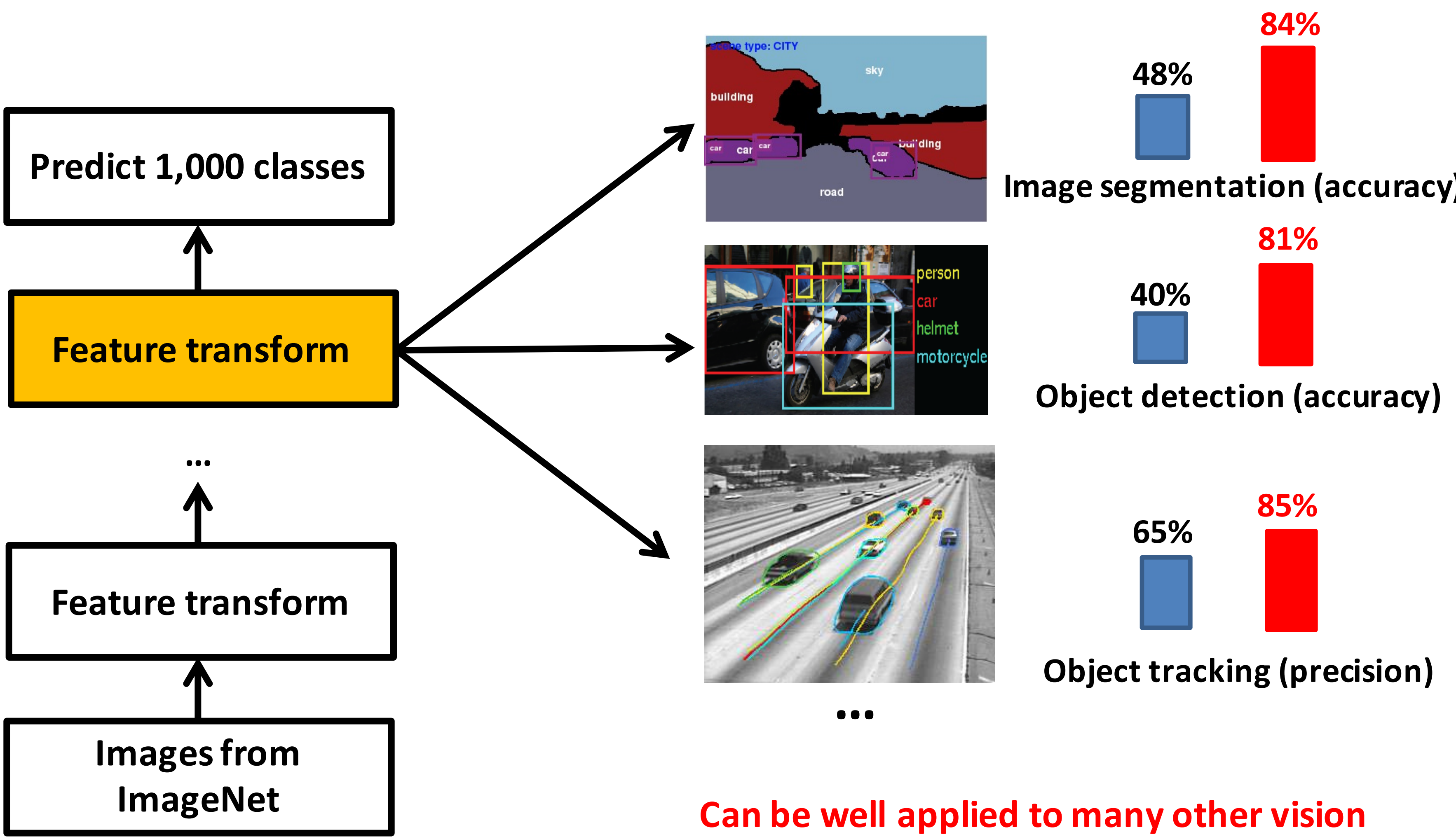
Images from ImageNet
will class labels



Learn feature
representations from
image classification task

How does human brain
learn about the world?

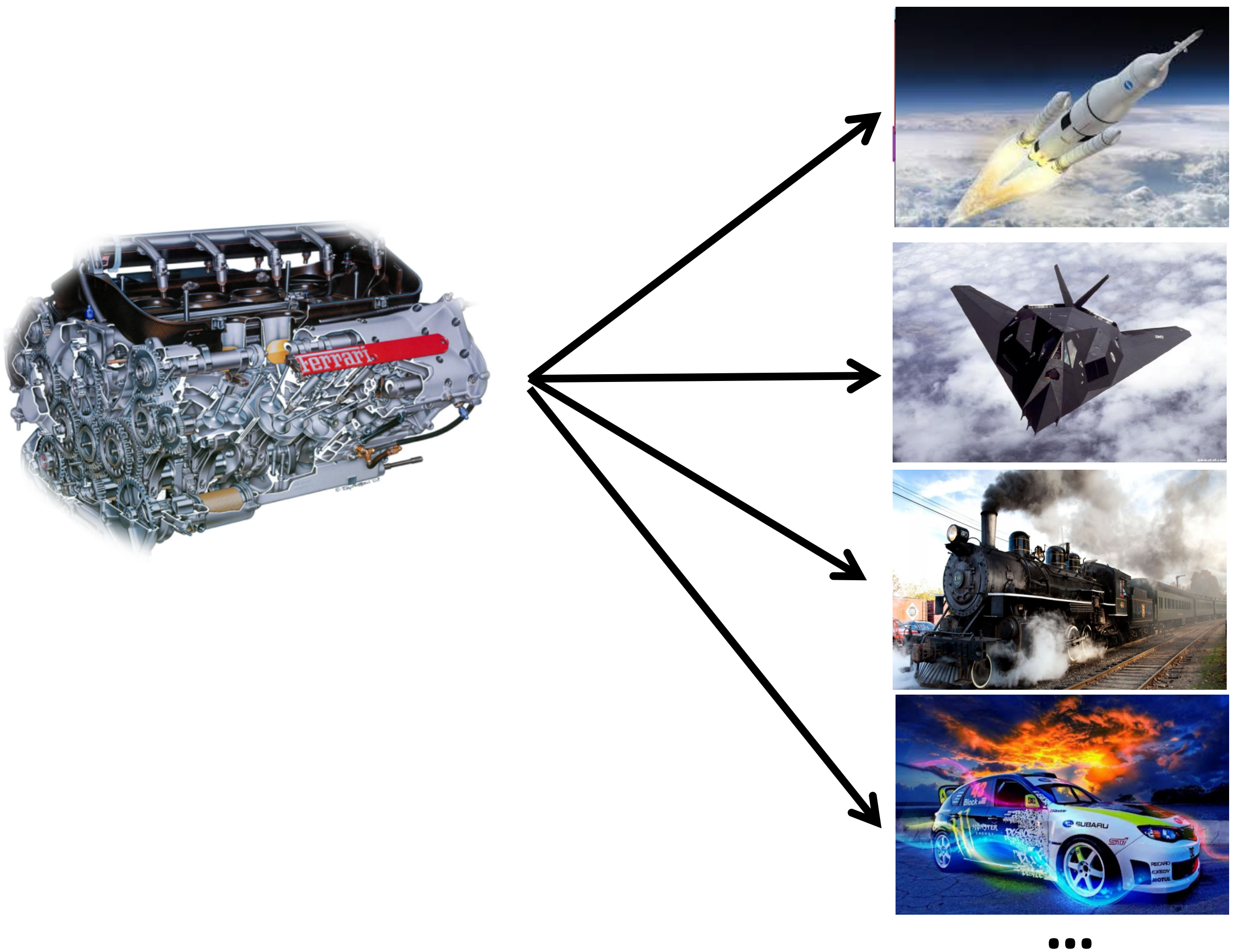
Deep Learning is a Universal Feature Learning Engine



Learning features from ImageNet

Can be well applied to many other vision tasks and datasets and boost their performance substantially

Deep Learning is a Universal Feature Learning Engine



Features learned from ImageNet serve as the engine driving many vision problems

How to increase model capacity?

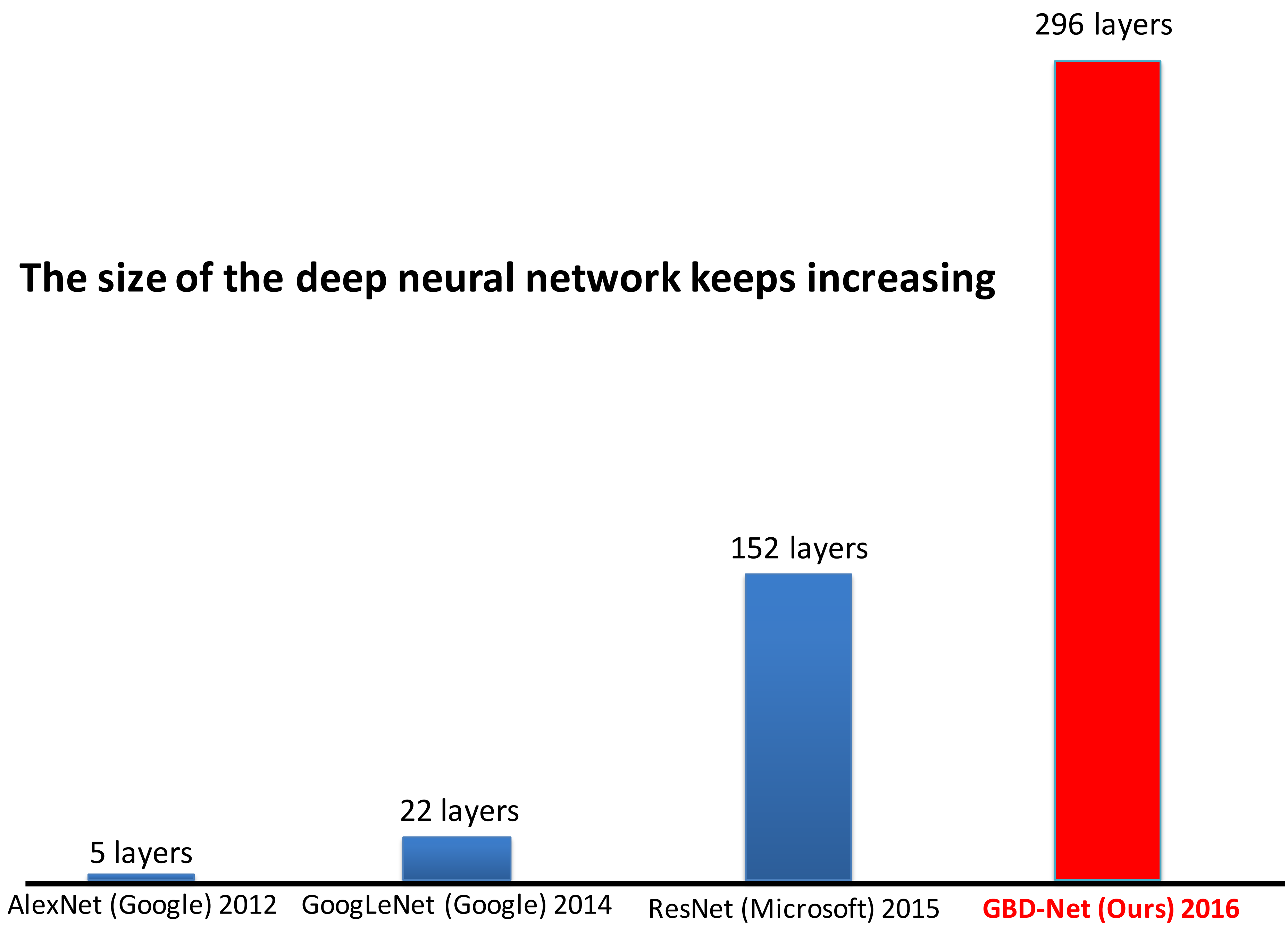
Curse of dimensionality

⇩

Blessing of dimensionality

⇩

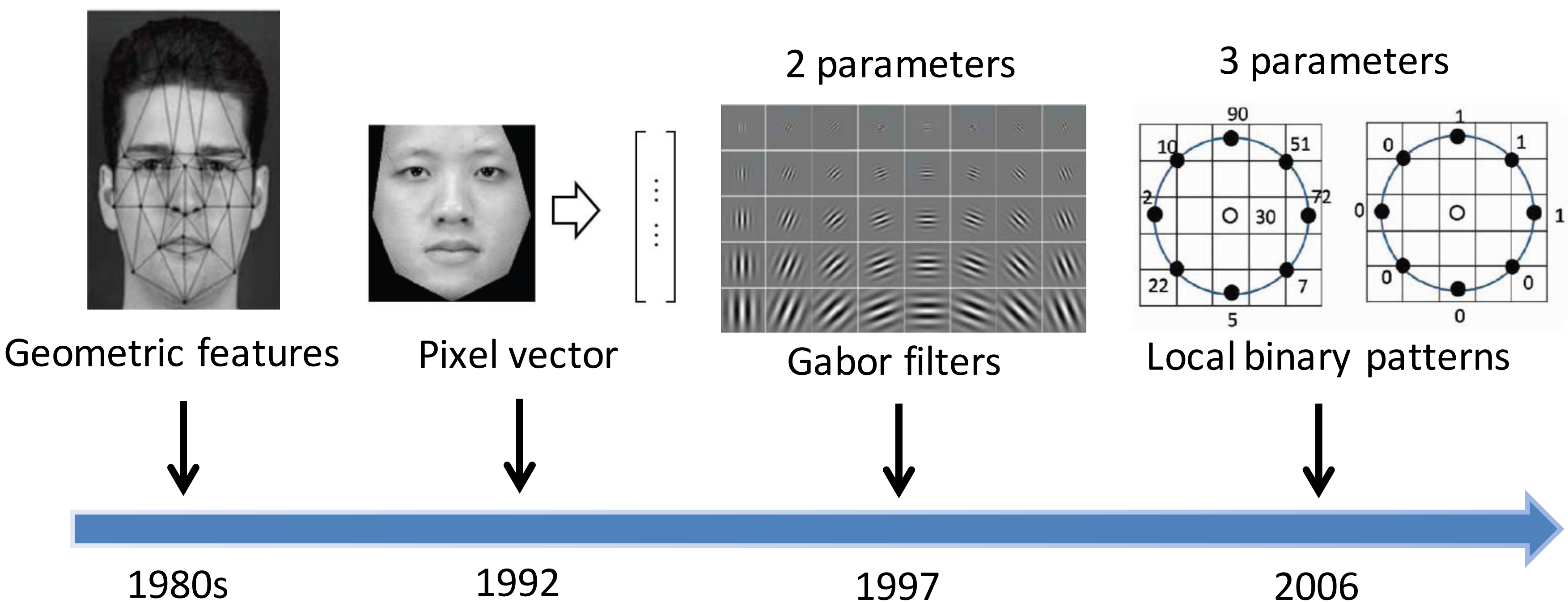
Learning hierarchical feature transforms
(Learning features with deep structures)


296 layers
The size of the deep neural network keeps increasing
152 layers
22 layers
5 layers
AlexNet (Google) 2012
GoogLeNet (Google) 2014
ResNet (Microsoft) 2015
GBD-Net (Ours) 2016

- The performance of a pattern recognition system heavily depends on feature representations

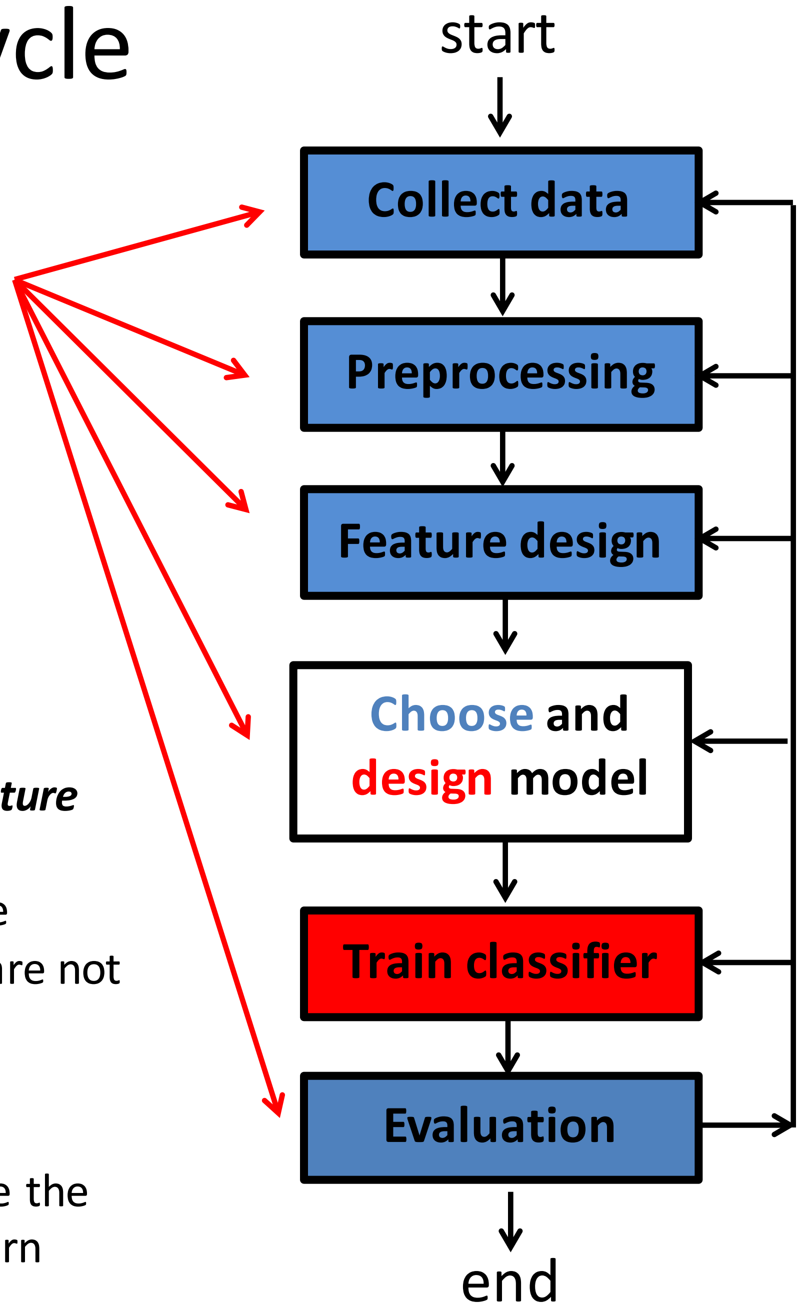
Feature engineering	Feature learning
Reply on human domain knowledge much more than data	Make better use of big data
If handcrafted features have multiple parameters, it is hard to manually tune them	Learn the values of a huge number of parameters in feature representations
Feature design is separate from training the classifier	Jointly learning feature transformations and classifiers makes their integration optimal
Developing effective features for new applications is slow	Faster to get feature representations for new applications

Handcrafted Features for Face Recognition



Design Cycle

Domain knowledge



Preprocessing and ***feature design*** may lose useful information and not be optimized, since they are not parts of an end-to-end learning system

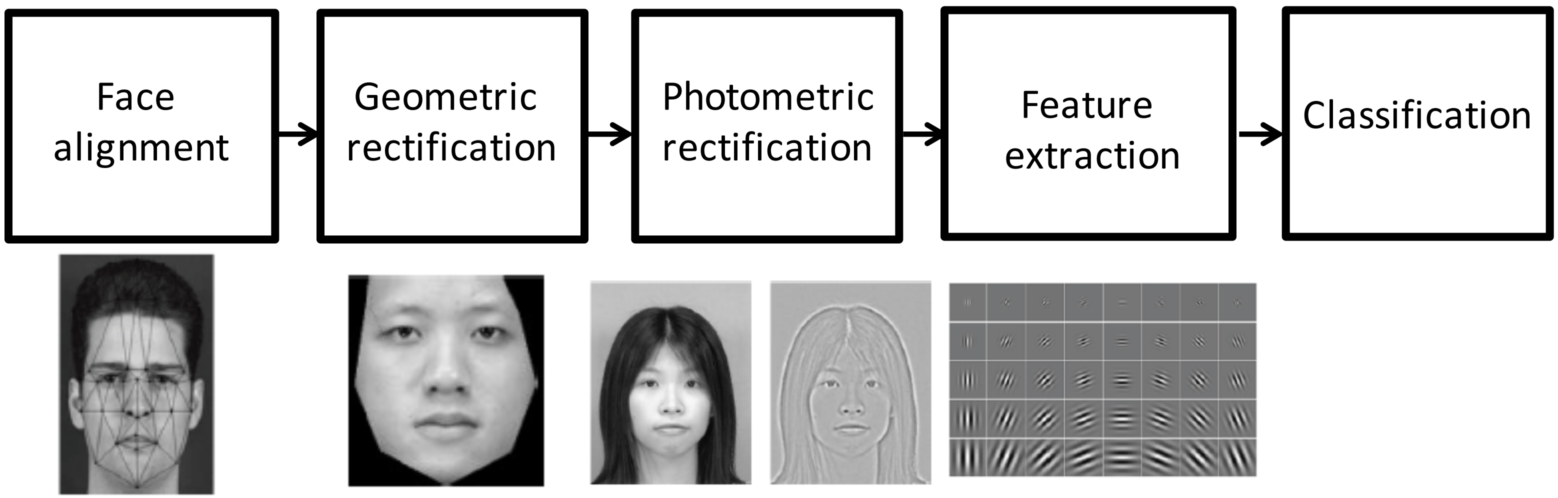
Preprocessing could be the result of another pattern recognition system

Interest of people working on **computer vision, speech recognition, medical image processing,…**

Interest of people working on **machine learning**

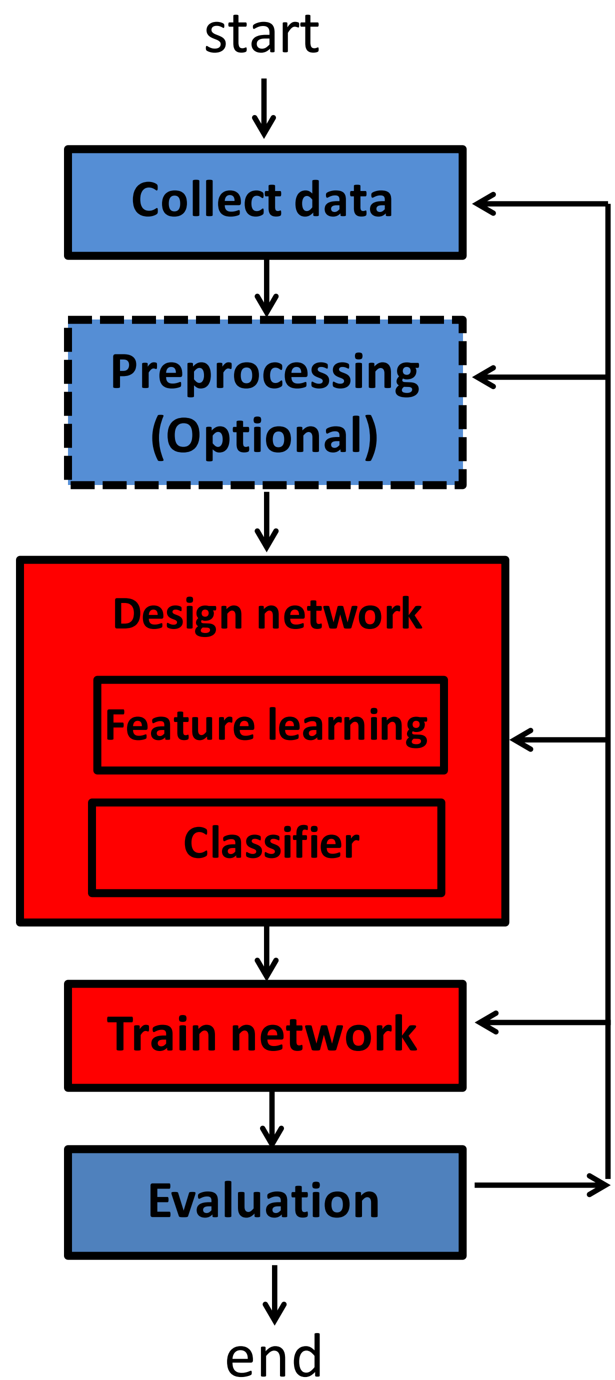
Interest of people working on **machine learning** and **computer vision, speech recognition, medical image processing,…**

Face recognition pipeline



Design Cycle with Deep Learning

- Learning plays a bigger role in the design cycle
- Feature learning becomes part of the end-to-end learning system
- Preprocessing becomes optional means that several pattern recognition steps can be merged into one end-to-end learning system
- Feature learning makes the key difference
- We underestimated the importance of data collection and evaluation



What makes deep learning successful in computer vision?

Li Fei-Fei

IMAGENET

Geoffrey Hinton

Data collection

Evaluation task

Deep learning

One million images with labels

Predict 1,000 image categories

CNN is not new

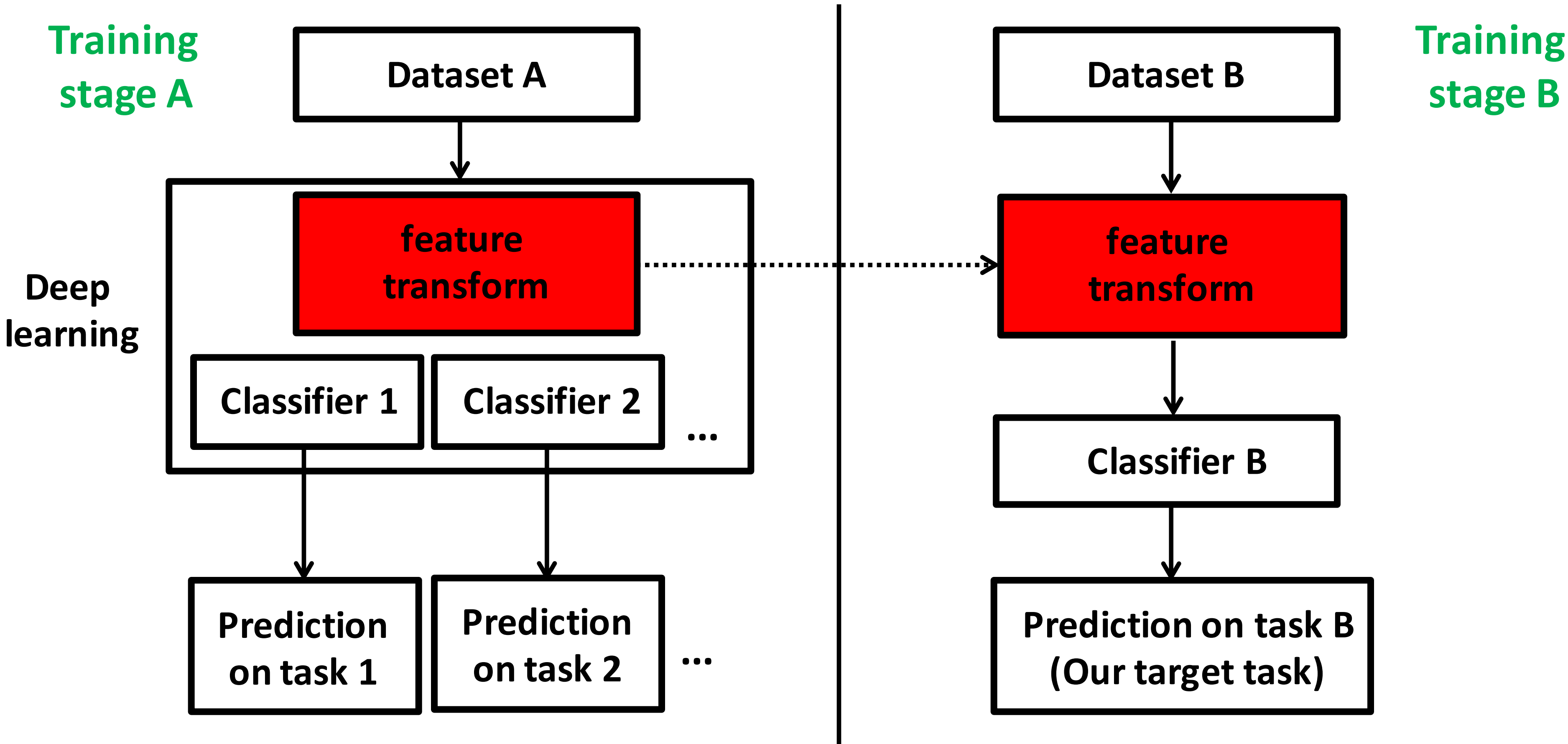
Design network structure

New training strategies

Feature learned from ImageNet can be well generalized to other tasks and datasets!

Learning features and classifiers separately

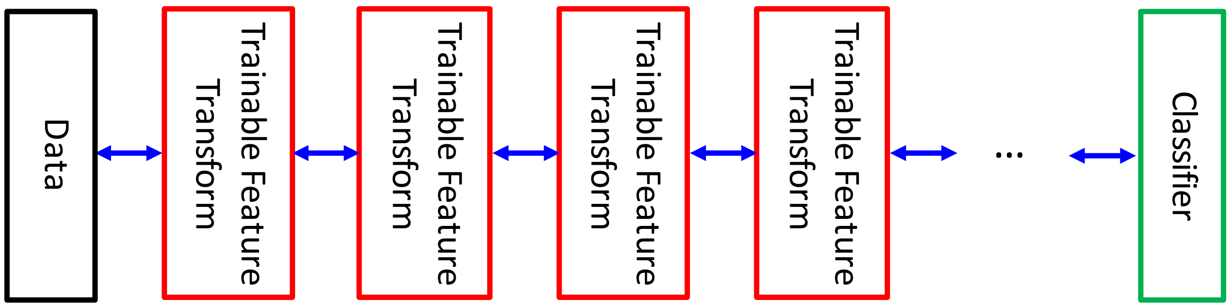
- Not all the datasets and prediction tasks are suitable for learning features with deep models



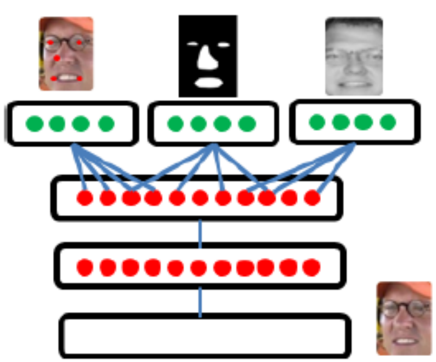
Deep Learning Means Feature Learning

- Deep learning is about learning hierarchical feature representations

$$\mathbf{y} = F(\mathbf{W}^k \cdot F(\mathbf{W}^{k-1} \cdot F(\ldots F(\mathbf{W}^0 \cdot \mathbf{x}))$$



- Good feature representations should be able to disentangle multiple factors coupled in the data

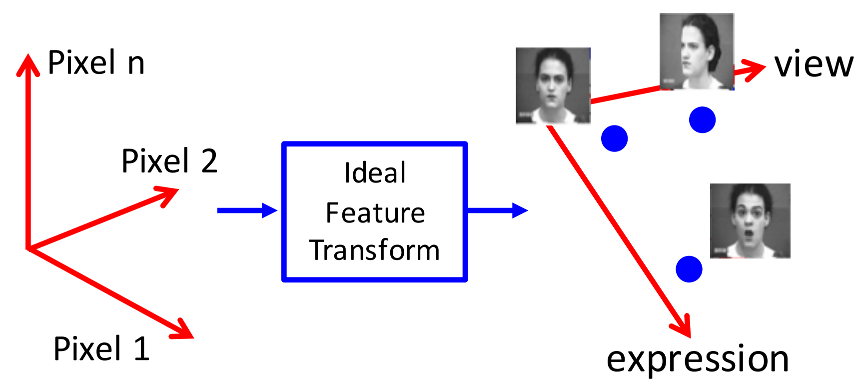


Identity: face recognition

Pose: pose estimation

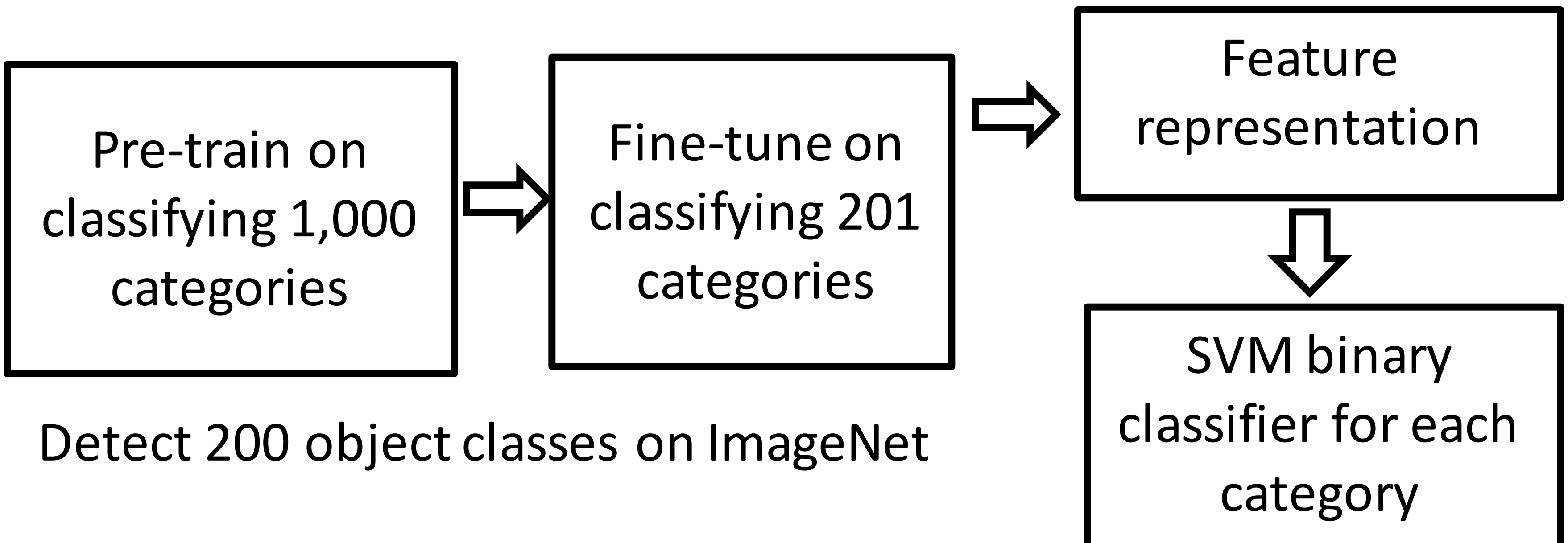
Expression: expression recognition

Age: age estimation



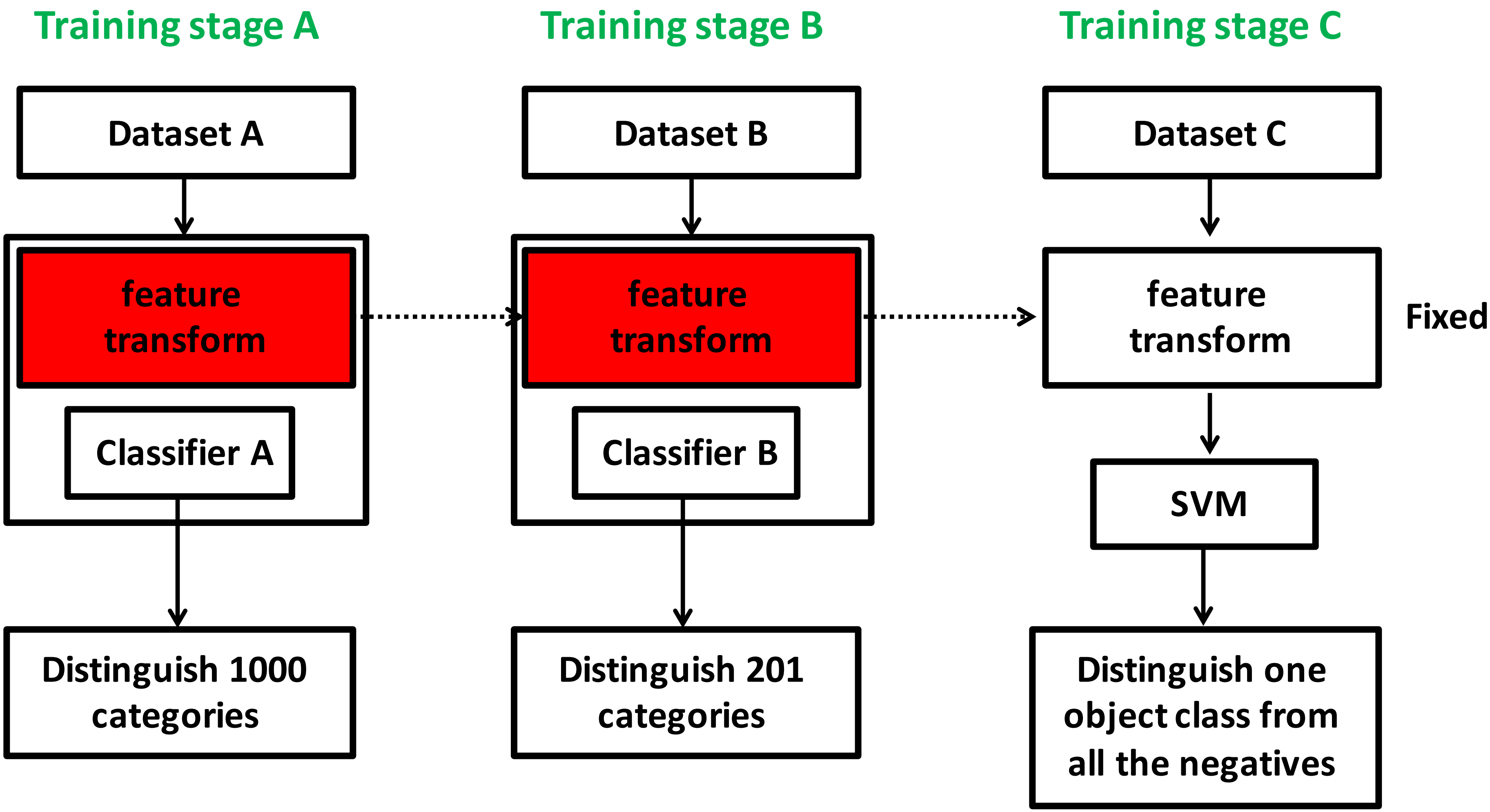
Example 1: General object detection on ImageNet

- How to effectively learn features with deep models
  - With challenging tasks
  - Predict high-dimensional vectors

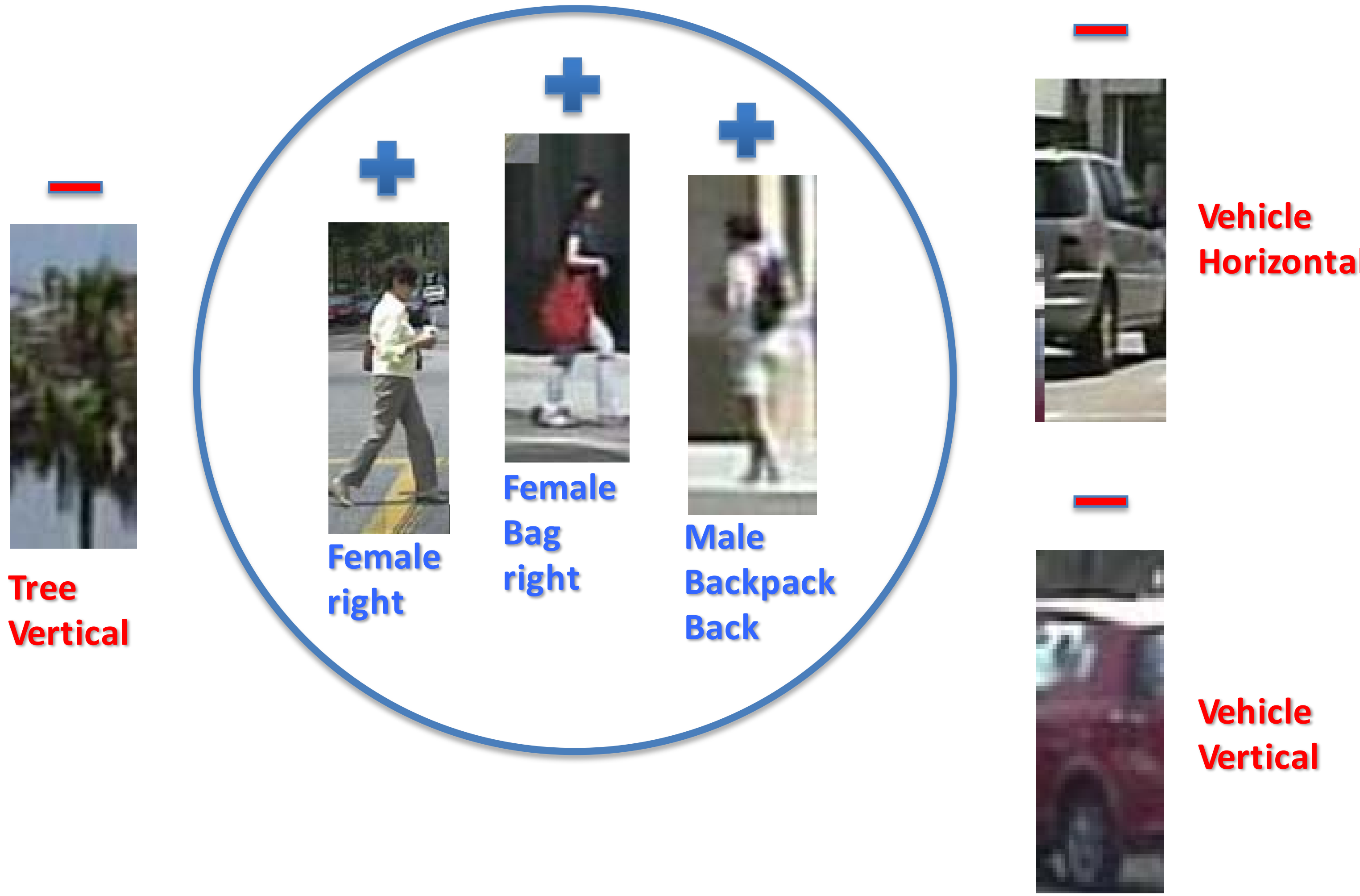


Detect 200 object classes on ImageNet

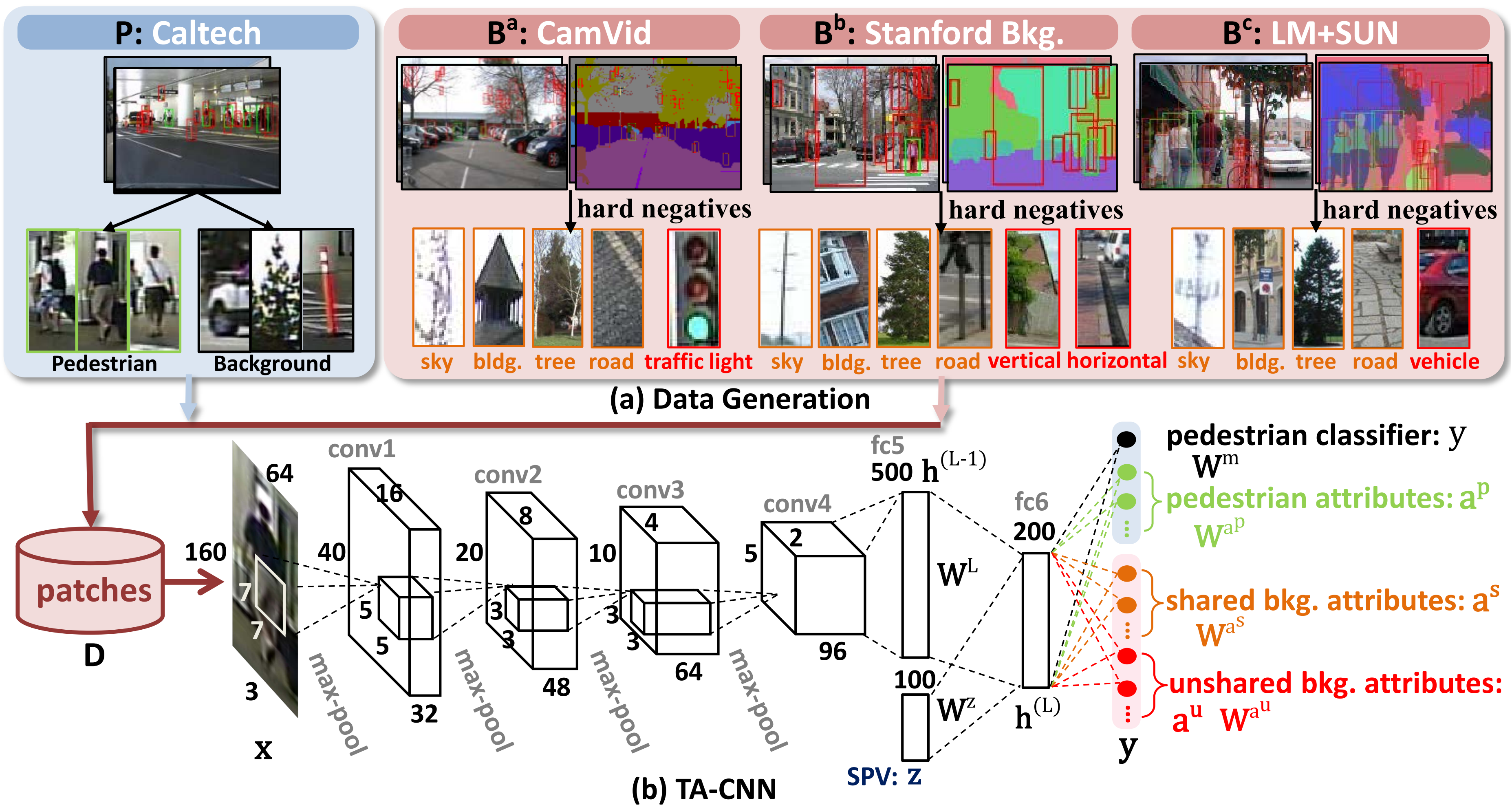
W. Ouyang and X. Wang et al. "DeepID-Net: deformable deep convolutional neural networks for object detection", CVPR, 2015


Training stage A
Training stage B
Training stage C
Dataset A
Dataset B
Dataset C
feature transform
feature transform
feature transform
Fixed
Classifier A
Classifier B
SVM
Distinguish 1000 categories
Distinguish 201 categories
Distinguish one object class from all the negatives

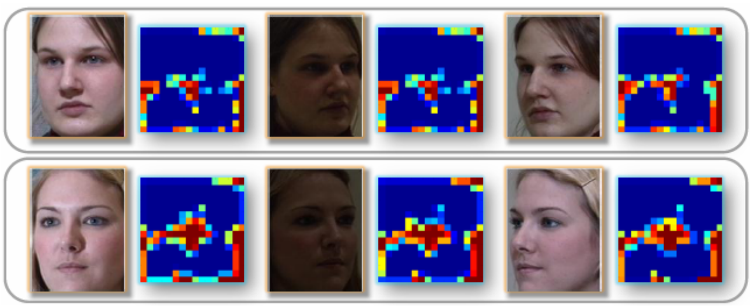
Example 2: Pedestrian detection aided by deep learning semantic tasks

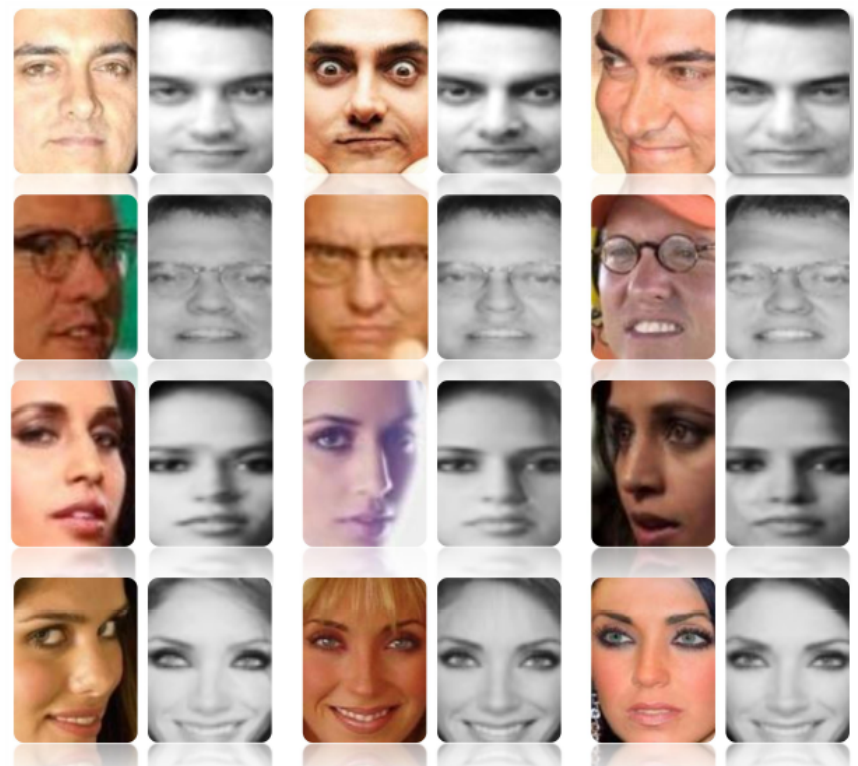


Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian Detection aided by Deep Learning Semantic Tasks," CVPR 2015


P: Caltech
Pedestrian
Background
Bᵃ: CamVid
Bᵇ: Stanford Bkg.
Bᶜ: LM+SUN
hard negatives
hard negatives
hard negatives
sky
bldg.
tree
road
traffic light
sky
bldg.
tree
road
vertical
horizontal
sky
bldg.
tree
road
vehicle
(a) Data Generation
patches
D
160
64
7
7
3
x
max-pool
conv1
16
40
5
5
32
max-pool
conv2
8
20
3
3
48
max-pool
conv3
4
10
3
3
64
max-pool
conv4
2
5
96
fc5
500 h(L-1)
W^L
100
W^z
SPV: z
fc6
200
h(L)
pedestrian classifier: y
W^m
pedestrian attributes: a^p
W^{a^p}
shared bkg. attributes: a^s
W^{a^s}
unshared bkg. attributes:
a^u W^{a^u}
y
(b) TA-CNN

Example 3: deep learning face identity features by recovering canonical-view face images

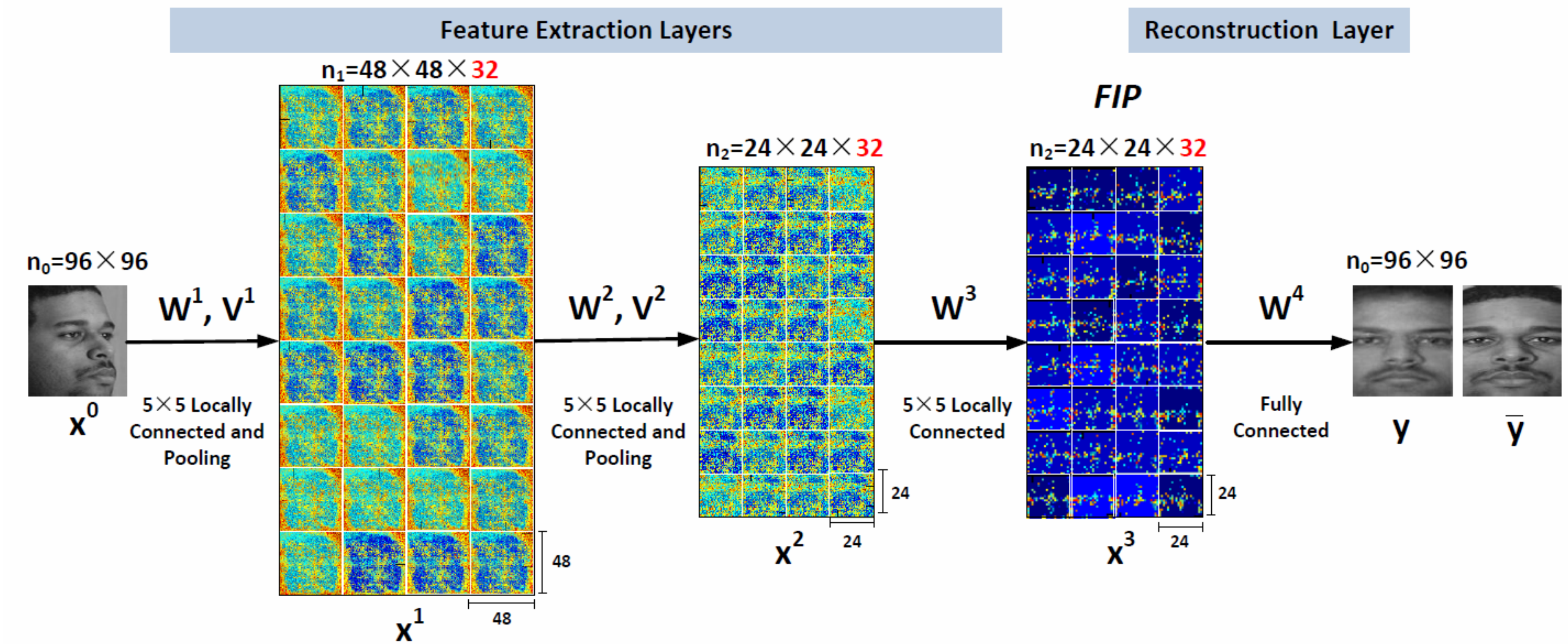




Reconstruction examples from LFW

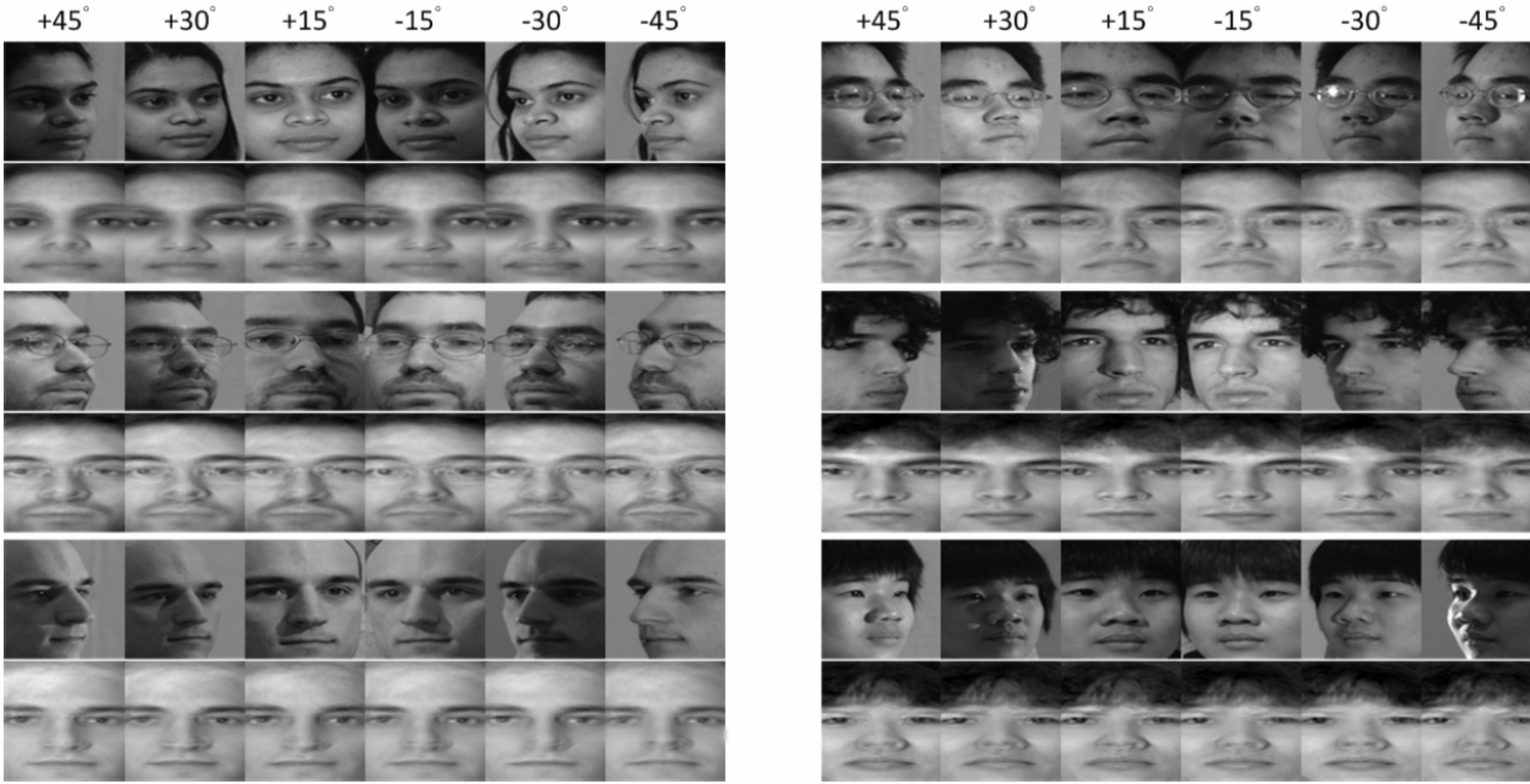
Z. Zhu, P. Luo, X. Wang, and X. Tang, “Deep Learning Identity Preserving Face Space,” ICCV 2013.

- Deep model can disentangle hidden factors through feature extraction over multiple layers
- No 3D model; no prior information on pose and lighting condition
- Model multiple complex transforms
- Reconstructing the whole face is a much strong supervision than predicting 0/1 class label and helps to avoid overfitting



Arbitrary view

Canonical view


+45°
+30°
+15°
-15°
-30°
-45°
+45°
+30°
+15°
-15°
-30°
-45°

Comparison on Multi-PIE

	-45°	-30°	-15°	+15°	+30°	+45°	Avg	Pose
LGBP [26]	37.7	62.5	77	83	59.2	36.1	59.3	√
VAAM [17]	74.1	91	95.7	95.7	89.5	74.8	86.9	√
FA-EGFC[3]	84.7	95	99.3	99	92.9	85.2	92.7	x
SA-EGFC[3]	93	**98.7**	99.7	**99.7**	**98.3**	93.6	97.2	√
LE[4] + LDA	86.9	95.5	99.9	**99.7**	95.5	81.8	93.2	x
CRBM[9] + LDA	80.3	90.5	94.9	96.4	88.3	89.8	87.6	x
Ours	**95.6**	**98.5**	**100.0**	**99.3**	**98.5**	**97.8**	**98.3**	x

[3] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith. Fully automatic pose-invariant face recognition via 3d pose normalization. In *ICCV*, pages 937–944, 2011. 1, 5, 6

[4] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *CVPR*, pages 2707–2714, 2010. 2, 3, 6

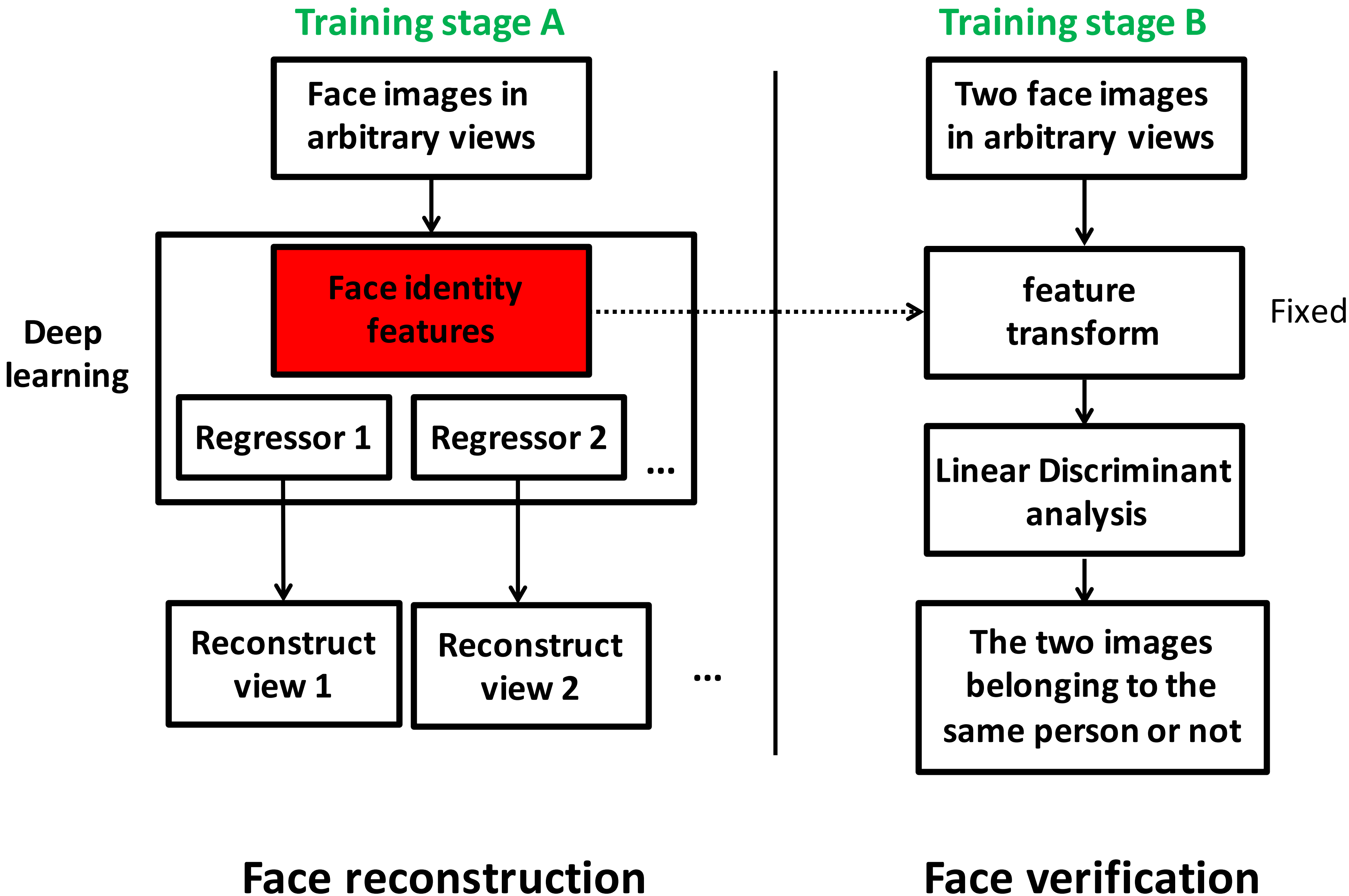
[9] G. B. Huang, H. Lee, and E. Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *CVPR*, pages 2518–2525, 2012. 3, 6

[17] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, and S. Shan. Morphable displacement field based image matching for face recognition across pose. In *ECCV*, pages 102–115. 2012. 1, 2, 5, 6

[26] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang. Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition. In *ICCV*, volume 1, pages 786–791, 2005. 5, 6

Deep learning 3D model from 2D images, mimicking human brain activities

Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep Learning and Disentangling Face Representation by Multi-View Perception," NIPS 2014.

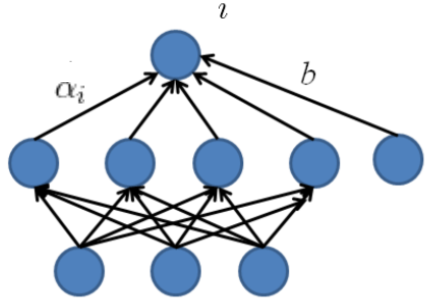

Training stage A
Face images in arbitrary views
Deep learning
Face identity features
Regressor 1
Regressor 2
...
Reconstruct view 1
Reconstruct view 2
...
Face reconstruction
Training stage B
Two face images in arbitrary views
feature transform
Fixed
Linear Discriminant analysis
The two images belonging to the same person or not
Face verification

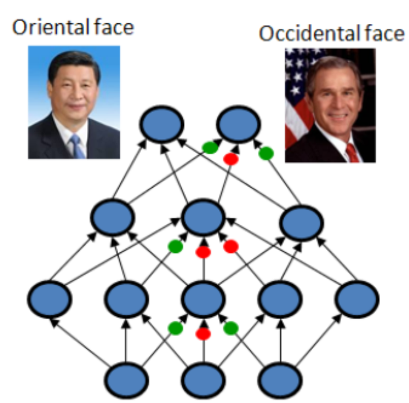
Deep Structures vs Shallow Structures
(Why deep?)

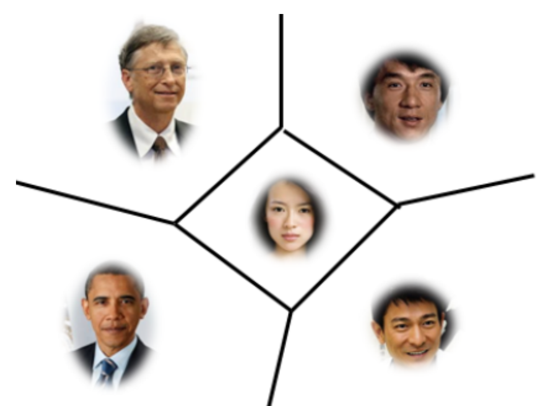
Shallow Structures

- A three-layer neural network (with one hidden layer) can approximate any classification function
- Most machine learning tools (such as SVM, boosting, and KNN) can be approximated as neural networks with one or two hidden layers
- Shallow models divide the feature space into regions and match templates in local regions. O(N) parameters are needed to represent N regions

SVM $g(x) = b + \sum_i \alpha_i K(x, x_i)$







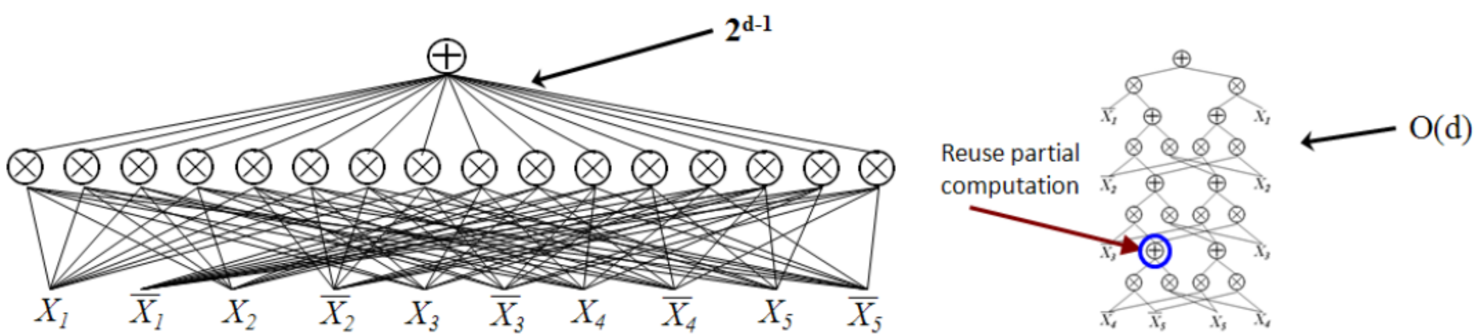
Deep Machines are More Efficient for Representing Certain Classes of Functions

- Theoretical results show that an architecture with insufficient depth can require many more computational elements, potentially exponentially more (with respect to input size), than architectures whose **depth is matched to the task** (Hastad 1986, Hastad and Goldmann 1991)
- It also means many more parameters to learn

- Take the d-bit parity function as an example

$$(X_1, \ldots, X_d) \in \{0,1\}^d \longmapsto \begin{cases} 1, & \text{if } \sum_{i=1}^{d} X_i \text{ is even} \\ -1, & \text{otherwise} \end{cases}$$

- d-bit logical parity circuits of depth 2 have exponential size (Andrew Yao, 1985)

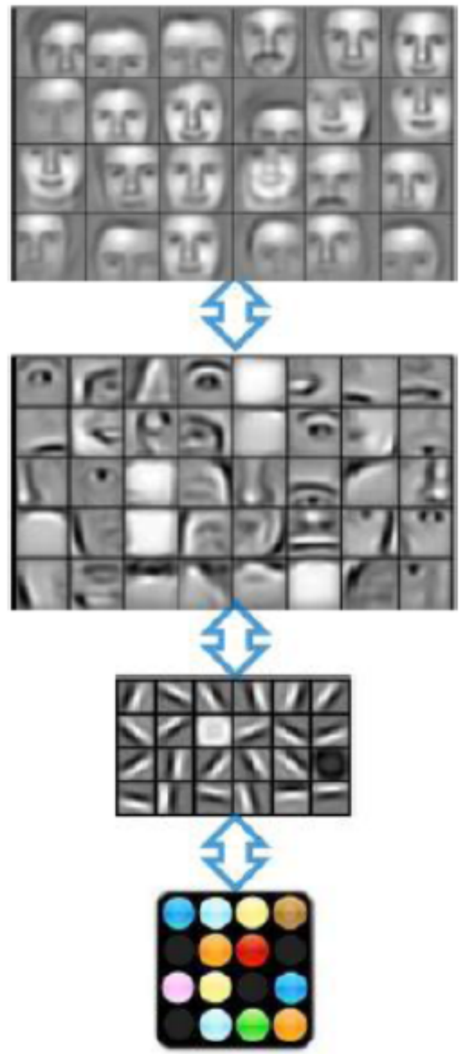


Shallow structure

Deep structure

- There are functions computable with a polynomial-size logic gates circuits of depth k that require exponential size when restricted to depth k -1 (Hastad, 1986)

- Architectures with multiple levels naturally provide sharing and re-use of components



Honglak Lee, NIPS’10

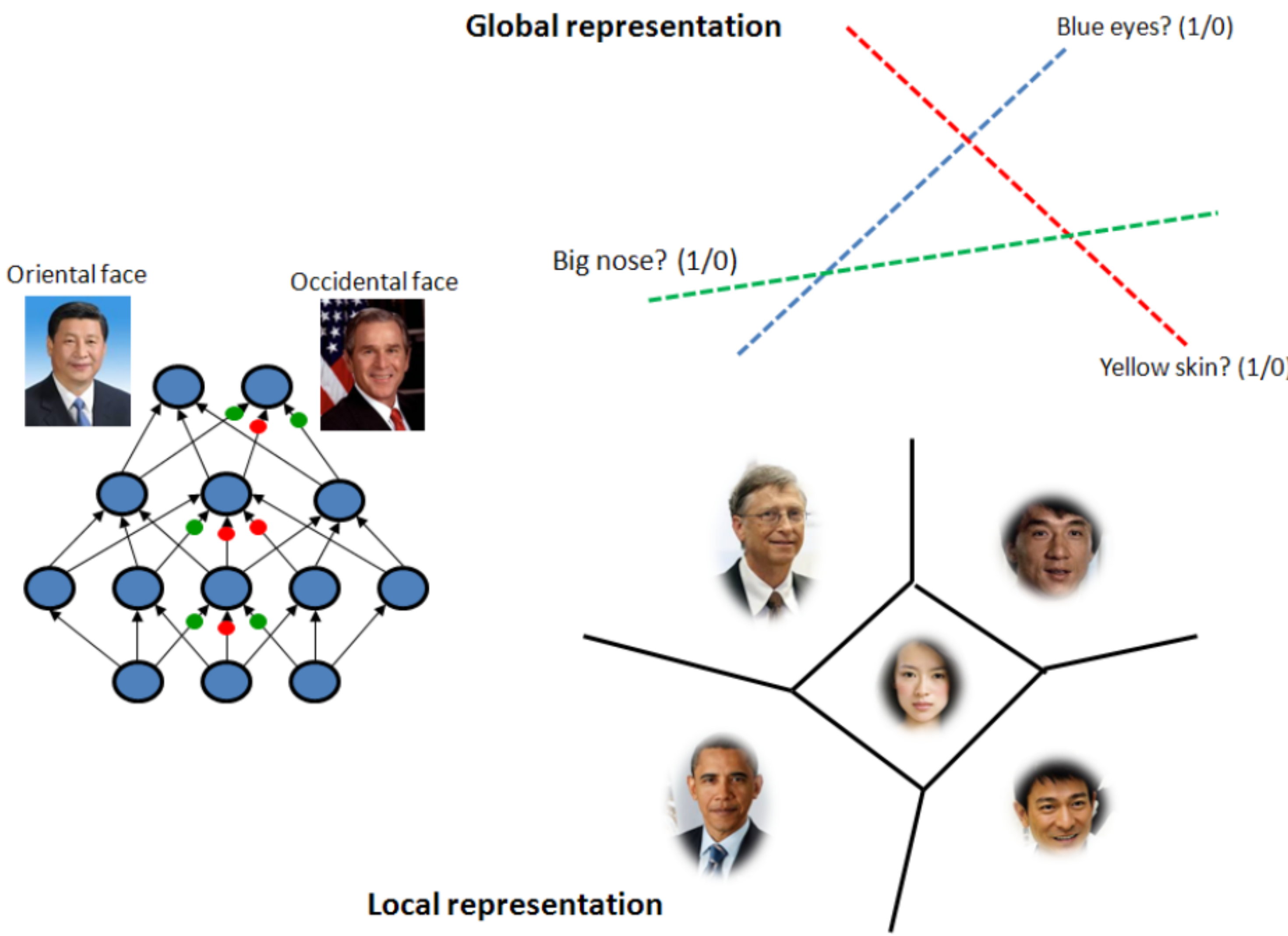
Humans Understand the World through Multiple Levels of Abstractions

- We do not interpret a scene image with pixels
  - Objects (sky, cars, roads, buildings, pedestrians) -> parts (wheels, doors, heads) -> texture -> edges -> pixels
  - Attributes: blue sky, red car
- It is natural for humans to decompose a complex problem into sub-problems through multiple levels of representations

Humans Understand the World through Multiple Levels of Abstractions

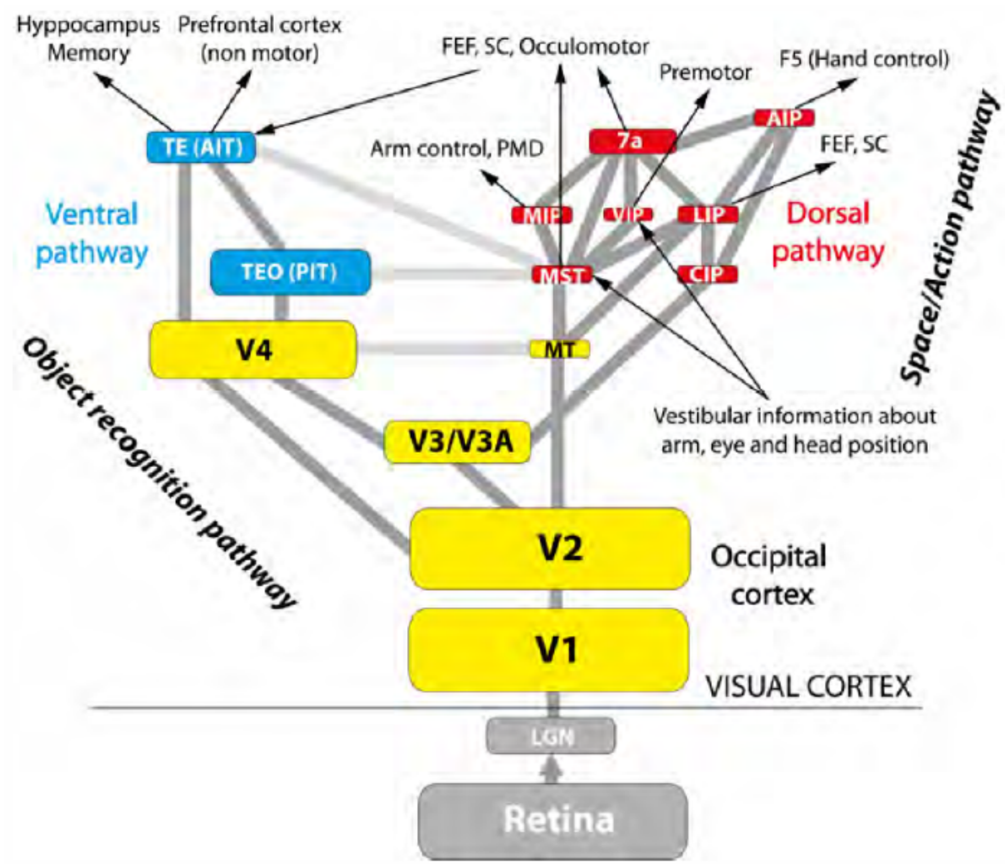
- Humans learn abstract concepts on top of less abstract ones
- Humans can imagine new pictures by re-configuring these abstractions at multiple levels. Thus our brain has good generalization can recognize things never seen before.
  - Our brain can estimate shape, lighting and pose from a face image and generate new images under various lightings and poses. That’s why we have good face recognition capability.

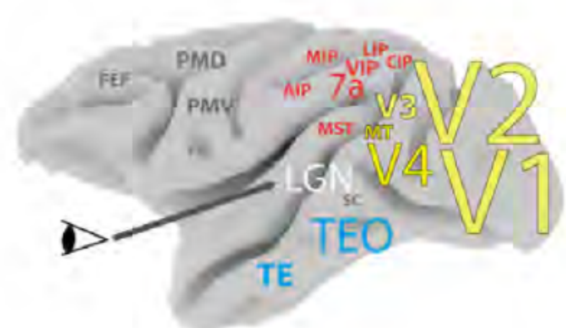
Local and Global Representations



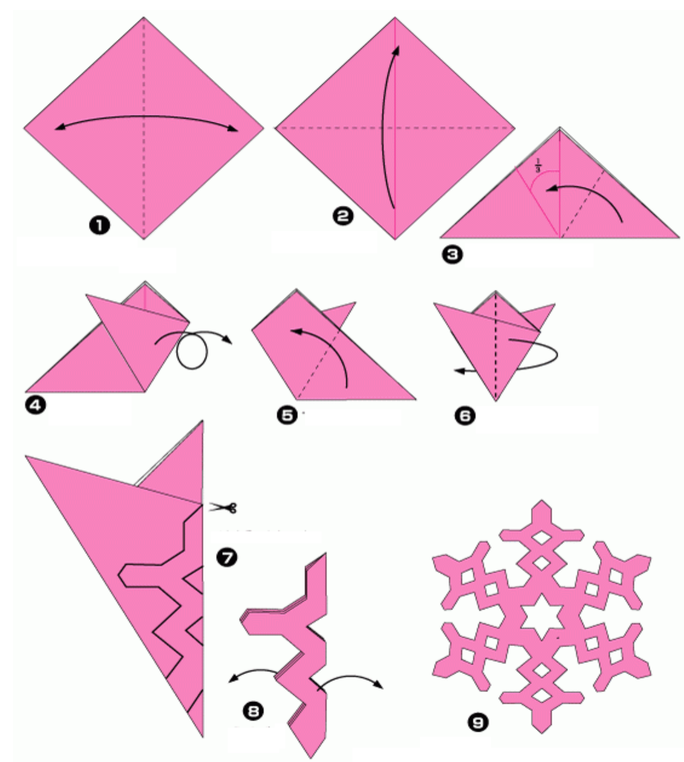
Human Brains Process Visual Signals through Multiple Layers

- A visual cortical area consists of six layers (Kruger et al. 2013)



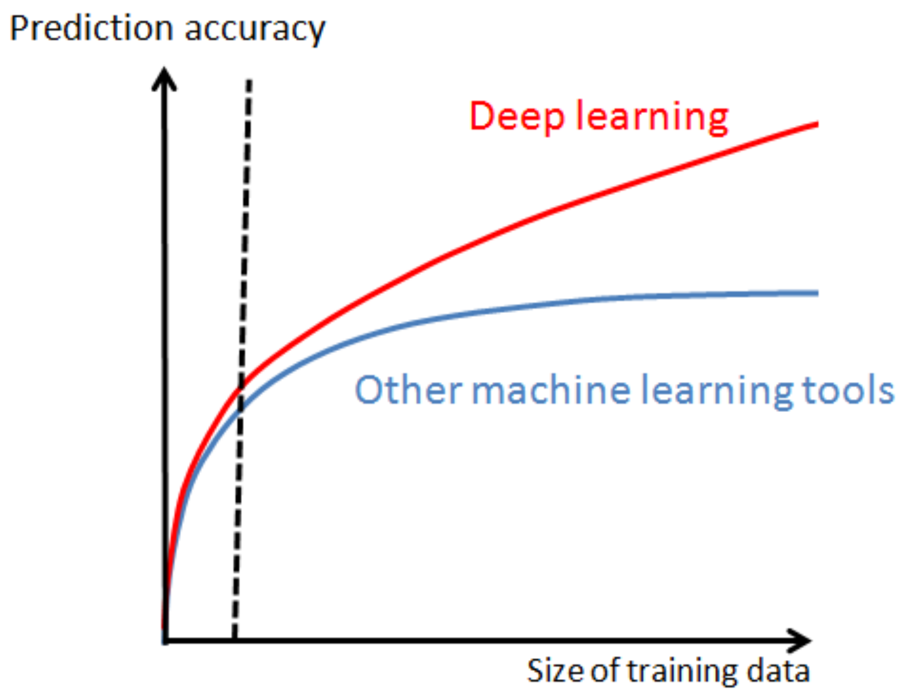


- The way these regions carve the input space still depends on few parameters: this huge number of regions are not placed independently of each other
- We can thus represent a function that looks complicated but actually has (global) structures



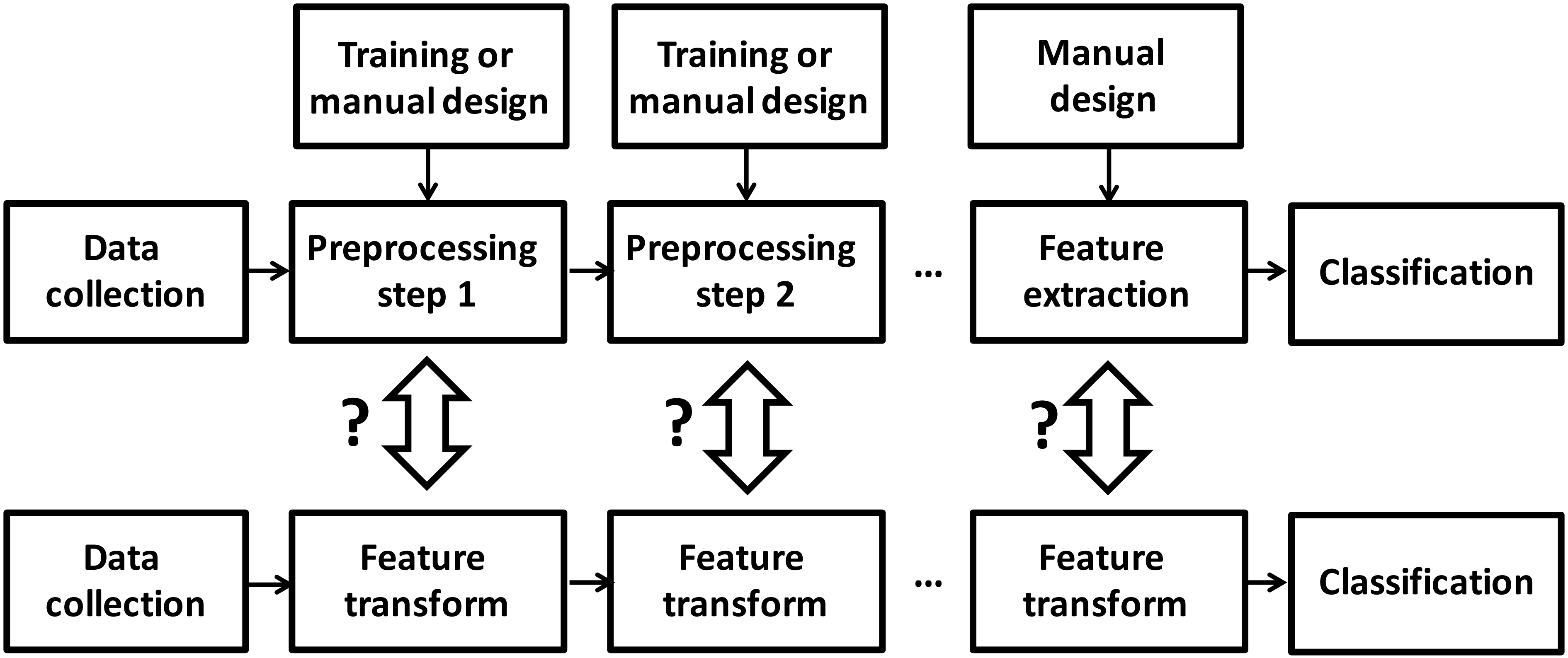
How do shallow models increase the model capacity?

- Typically increase the size of feature vectors



D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: Highdimensional feature and its efficient compression for face verification. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2013.

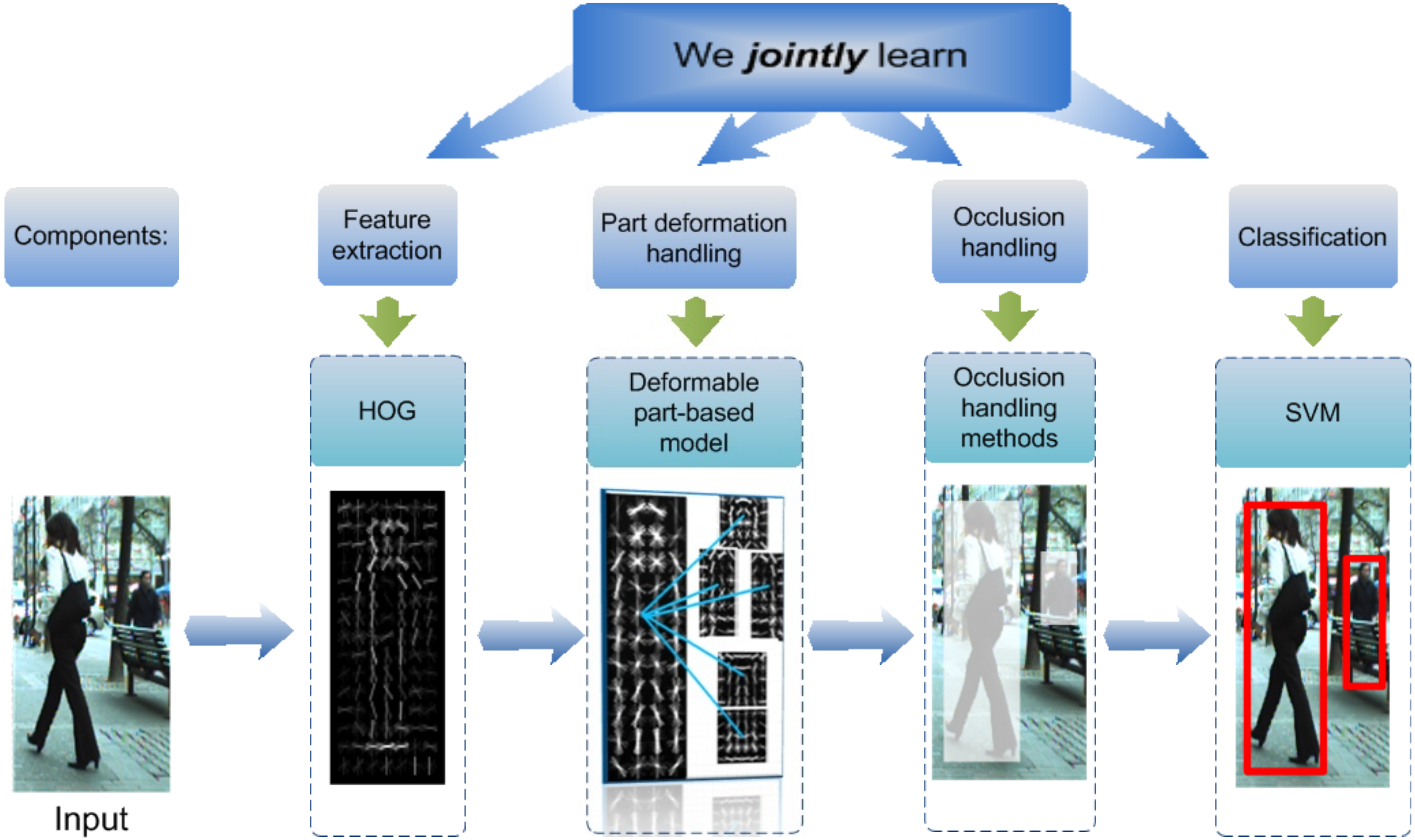
Joint Learning vs Separate Learning



End-to-end learning

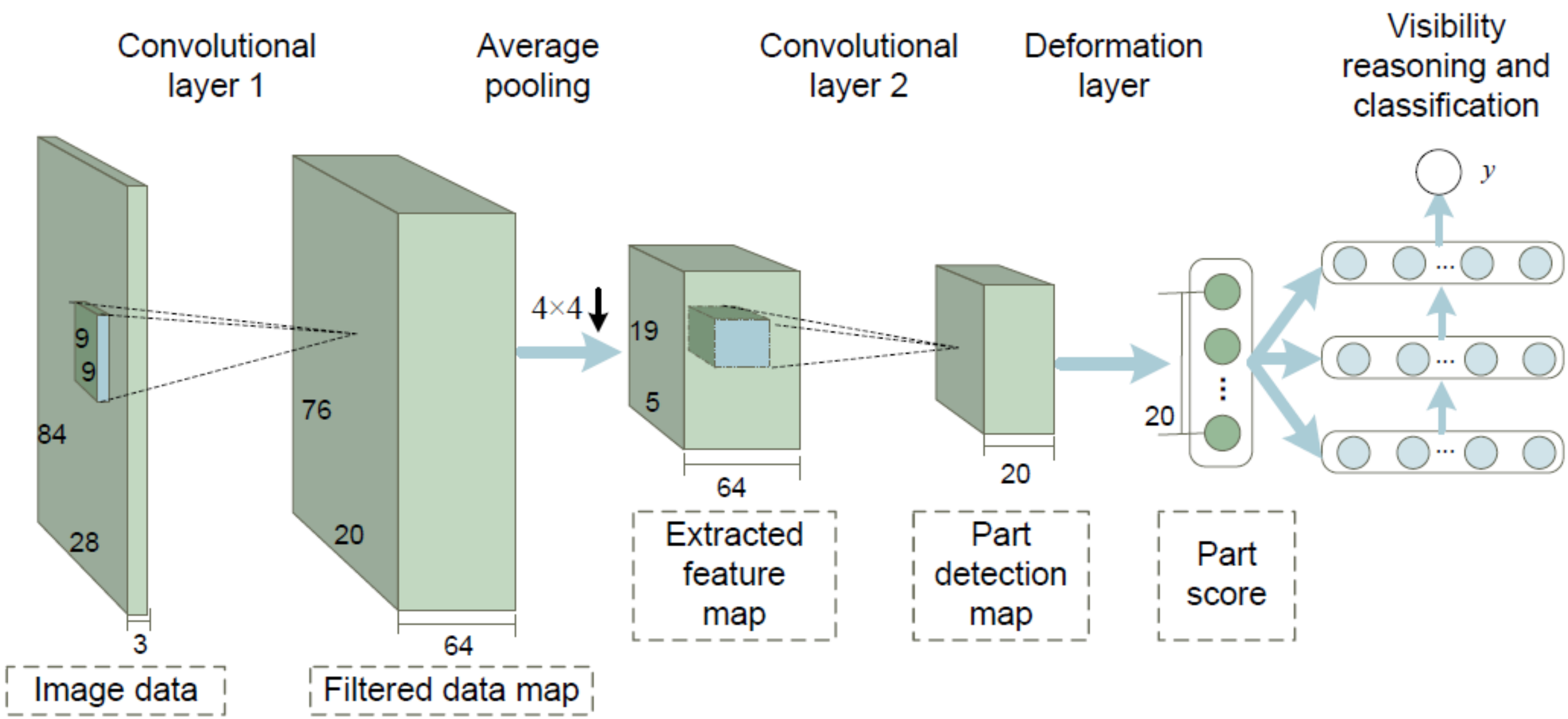
Deep learning is a framework/language but not a black-box model

Its power comes from joint optimization and increasing the capacity of the learner



- N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. CVPR, 2005. (6000 citations)
- P. Felzenszwalb, D. McAlester, and D. Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. CVPR, 2008. (2000 citations)
- W. Ouyang and X. Wang. A Discriminative Deep Model for Pedestrian Detection with Occlusion Handling. CVPR, 2012.

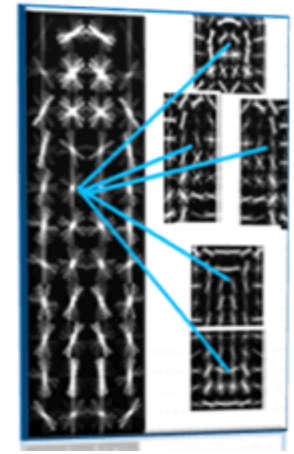
Our Joint Deep Learning Model



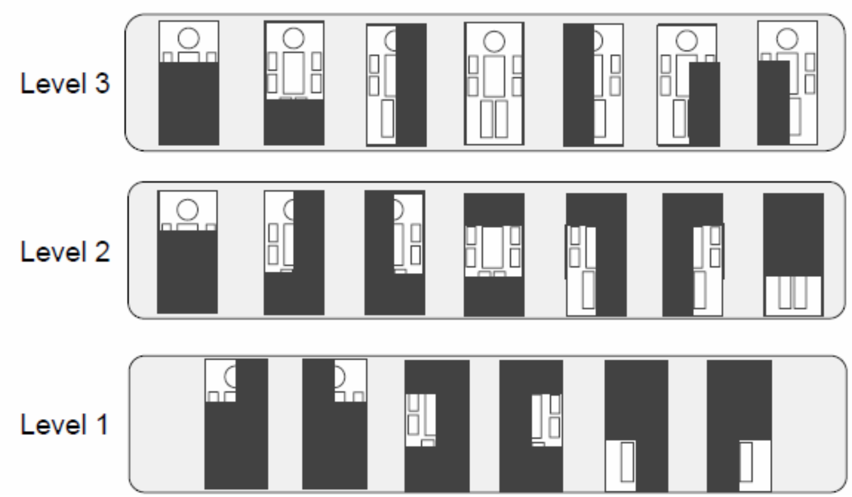
W. Ouyang and X. Wang, "Joint Deep Learning for Pedestrian Detection," Proc. ICCV, 2013.

Modeling Part Detectors

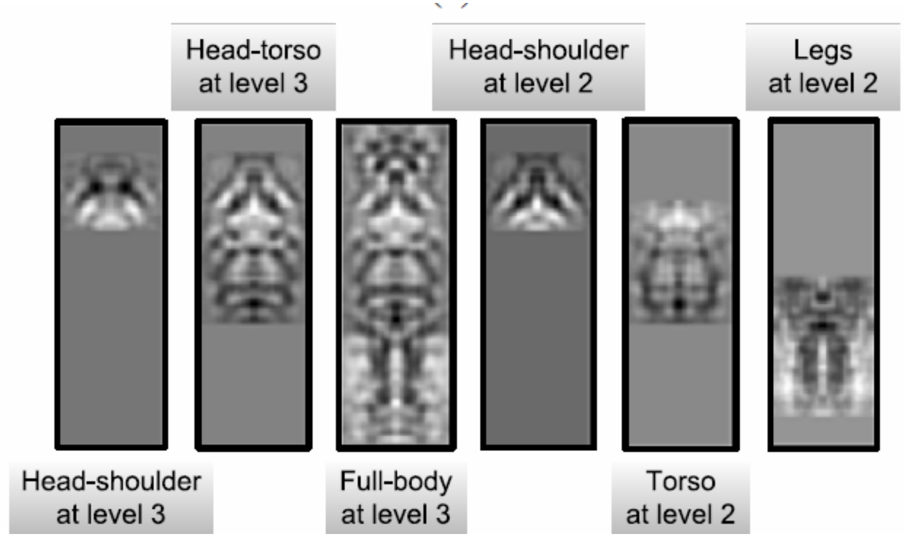
- Design the filters in the second convolutional layer with variable sizes



Part models learned from HOG

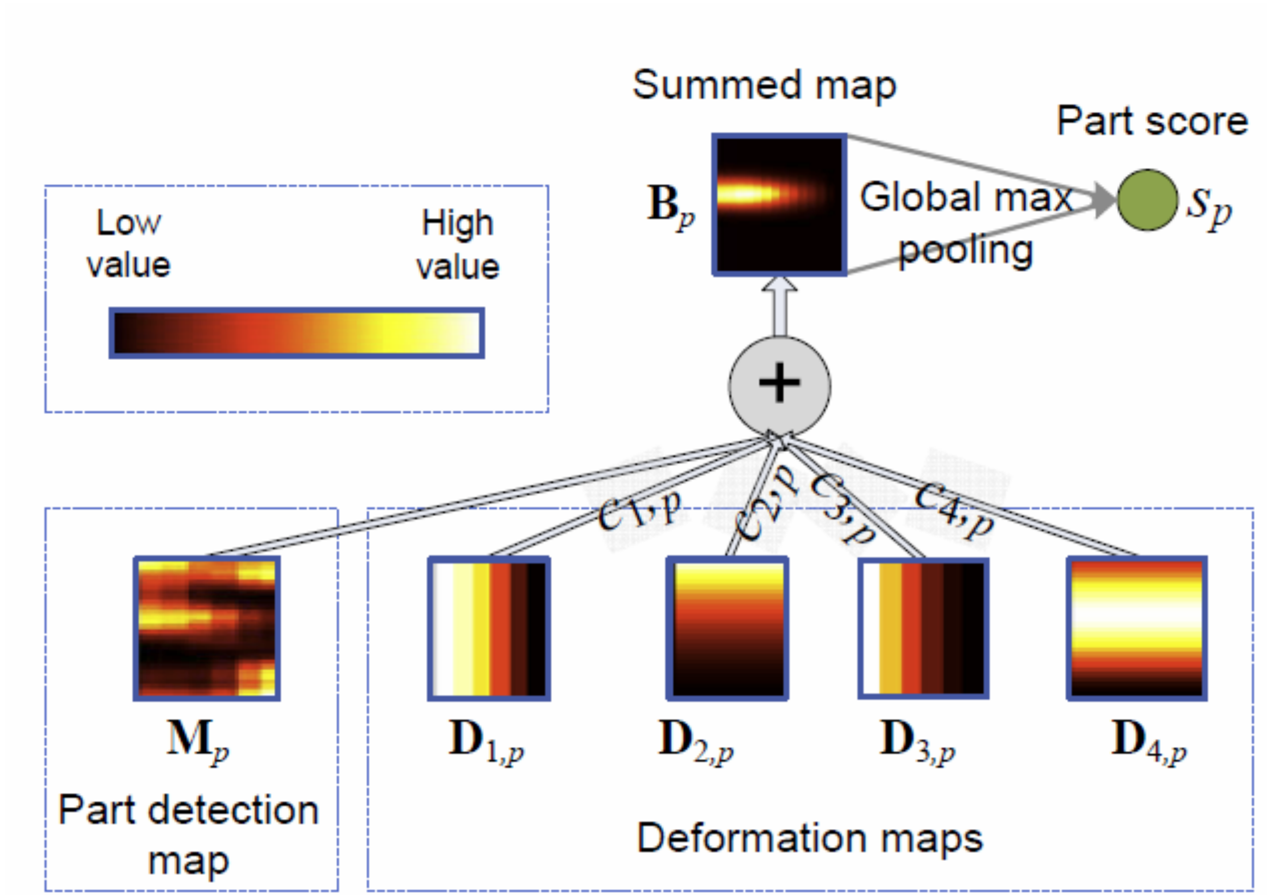


Part models

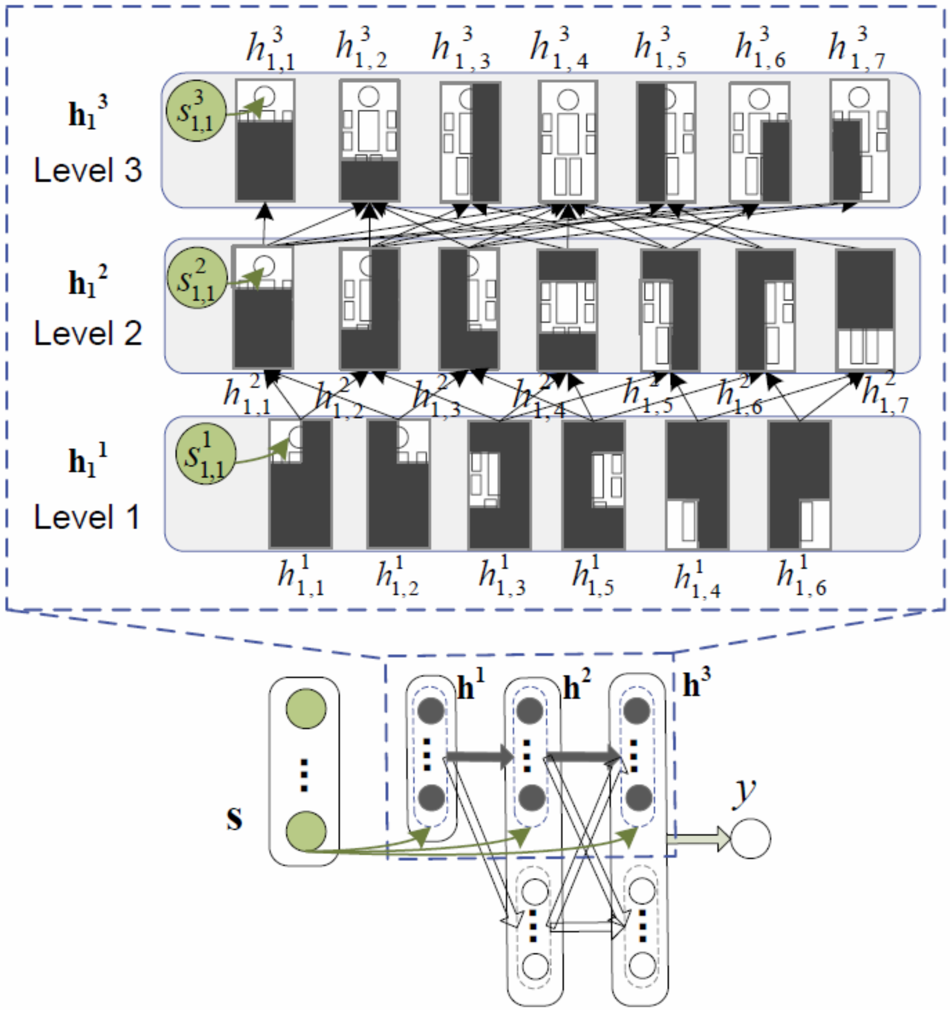


Learned filtered at the second convolutional layer

Deformation Layer



Visibility Reasoning with Deep Belief Net

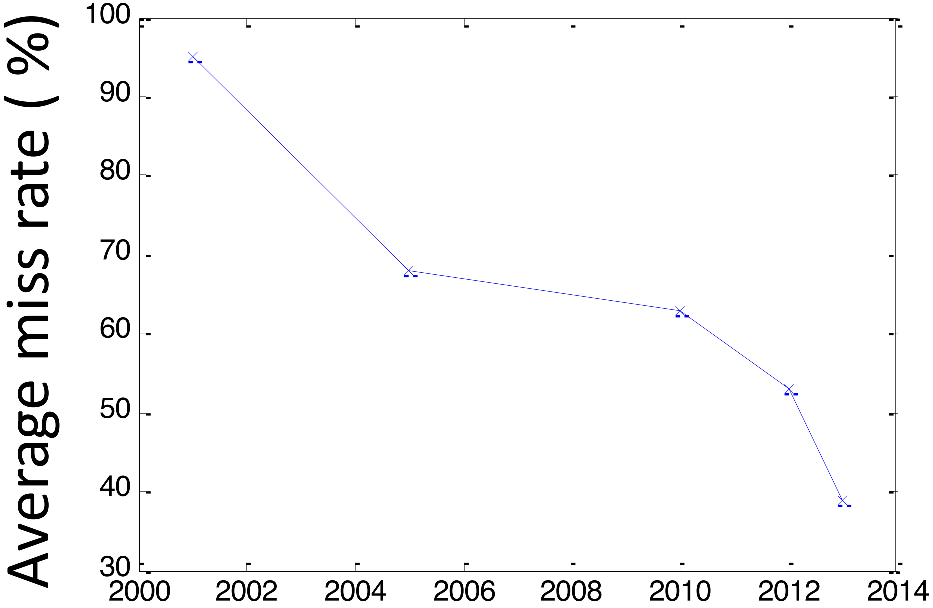


$$\tilde{h}_j^{l+1} = \sigma(\tilde{\mathbf{h}}^{l\mathrm{T}} \mathbf{w}_{*,j}^{l} + c_j^{l+1} + g_j^{l+1} s_j^{l+1})$$

Correlates with part detection score

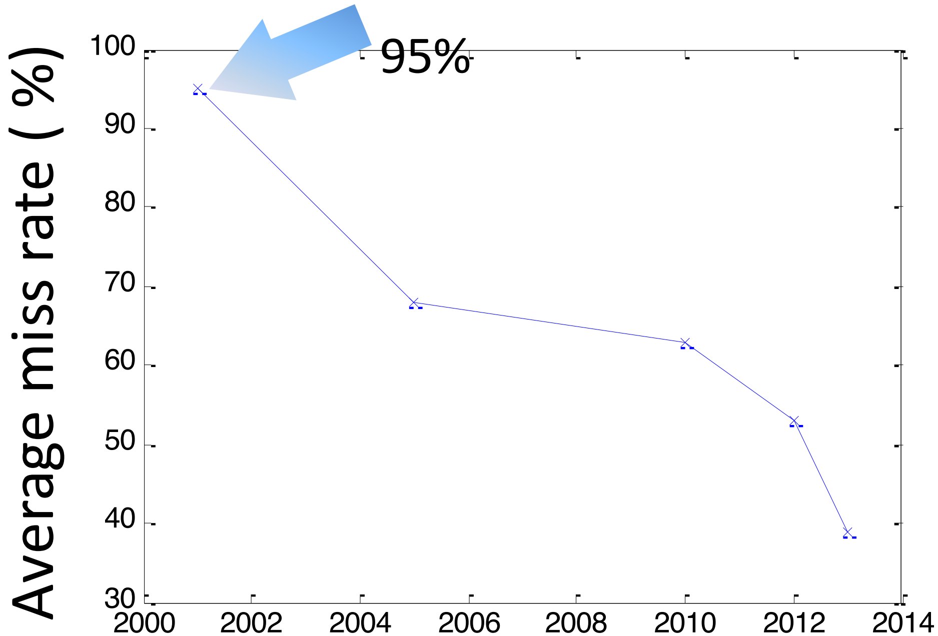
Experimental Results

- Caltech – Test dataset (largest, most widely used)



Experimental Results

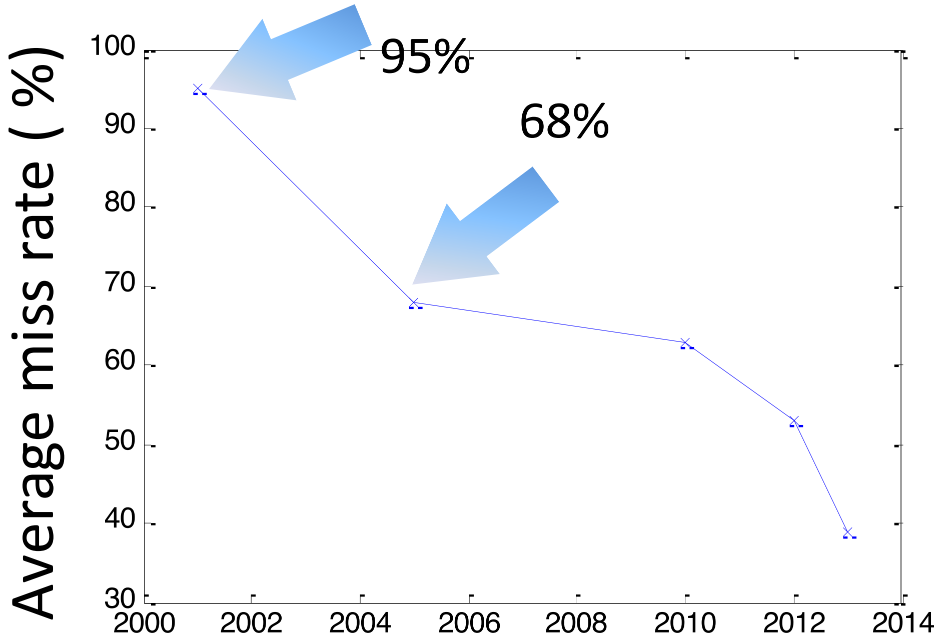
- Caltech – Test dataset (largest, most widely used)



Rapid object detection using a boosted cascade of simple features
P Viola, M Jones - … Vision and Pattern Recognition, 2001. CVPR …, 2001 - ieeexplore.ieee.org
Abstract This paper describes a machine learning approach for visual **object detection** which is capable of processing images extremely rapidly and achieving high **detection** rates. This work is distinguished by three key contributions. The first is the introduction of a new **…**
Cited by 7647 Related articles All 201 versions Import into BibTeX More

Experimental Results

- Caltech – Test dataset (largest, most widely used)



Histograms of oriented **gradients** for **human detection**

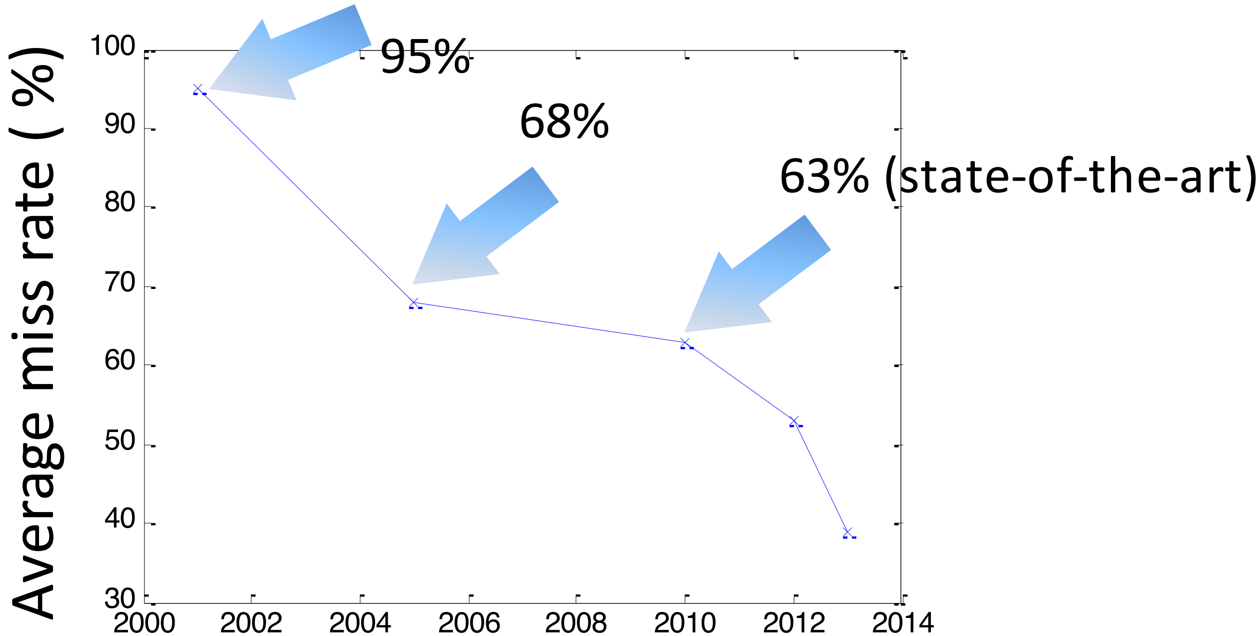
N Dalal, B Triggs - … and Pattern Recognition, 2005. CVPR 2005 …, 2005 - ieeexplore.ieee.org

… We study the issue of feature sets for **human detection**, showing that lo- cally normalized **Histogram** of Oriented **Gradient** (HOG) de- scriptors provide excellent performance relative to other ex- isting feature sets including wavelets [17,22]. **…**

Cited by 5438 Related articles All 106 versions Import into BibTeX More

Experimental Results

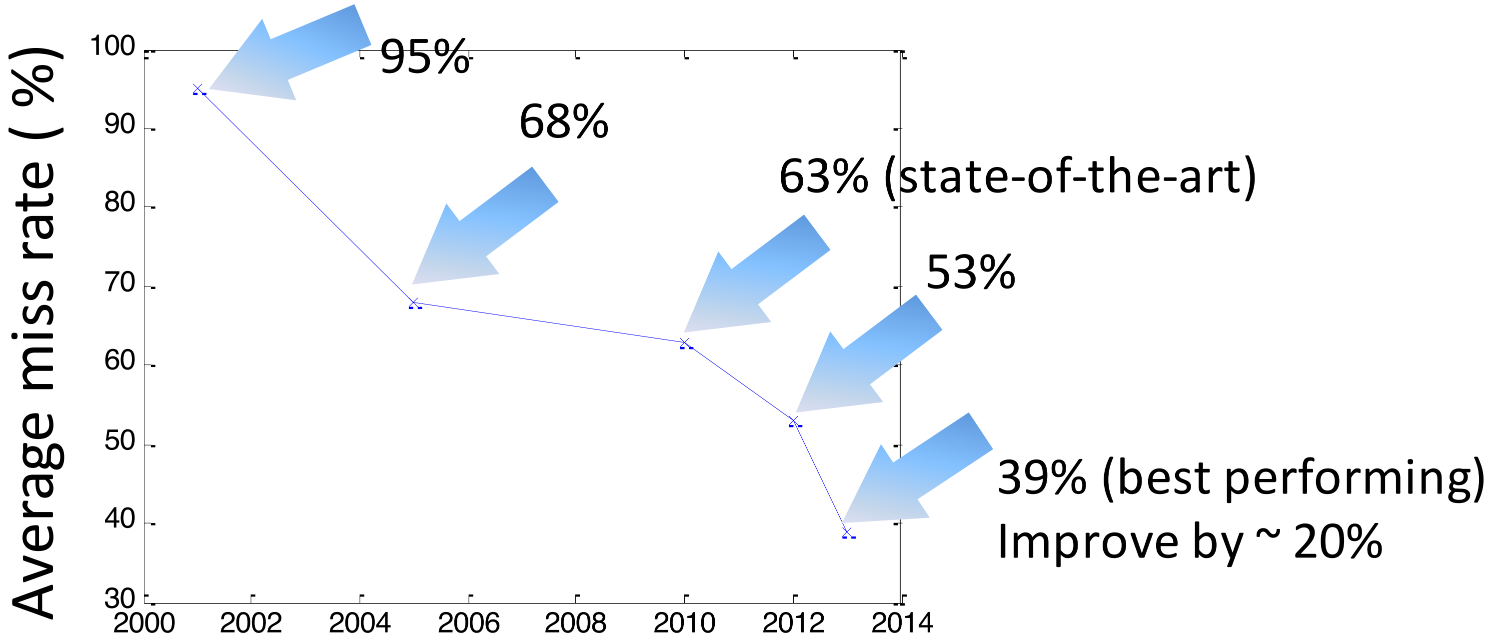
- Caltech – Test dataset (largest, most widely used)



Object detection with **discriminatively trained part-based models**
PF Felzenszwalb, RB Girshick… - Pattern Analysis and …, 2010 - ieeexplore.ieee.org
Abstract We describe an **object detection** system **based** on mixtures of multiscale deformable **part models**. Our system is able to represent highly variable **object** classes and achieves state-of-the-art results in the PASCAL **object detection** challenges. While **...**
Cited by 964   Related articles   All 43 versions   Import into BibTeX   More

Experimental Results

- Caltech – Test dataset (largest, most widely used)



W. Ouyang and X. Wang, "A Discriminative Deep Model for Pedestrian Detection with Occlusion Handling," CVPR 2012.

W. Ouyang, X. Zeng and X. Wang, "Modeling Mutual Visibility Relationship in Pedestrian Detection ", CVPR 2013.

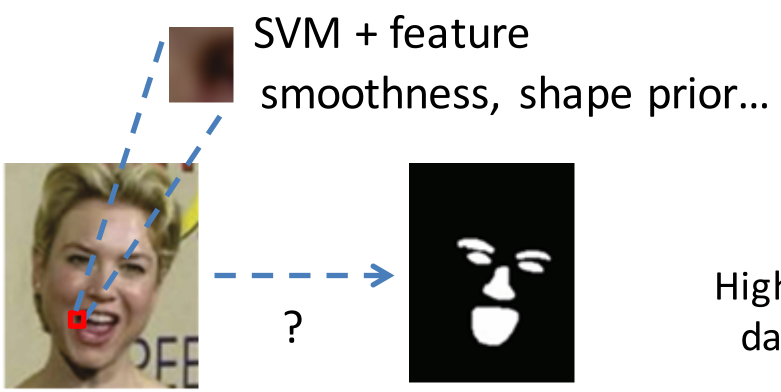
W. Ouyang, Xiaogang Wang, "Single-Pedestrian Detection aided by Multi-pedestrian Detection ", CVPR 2013.

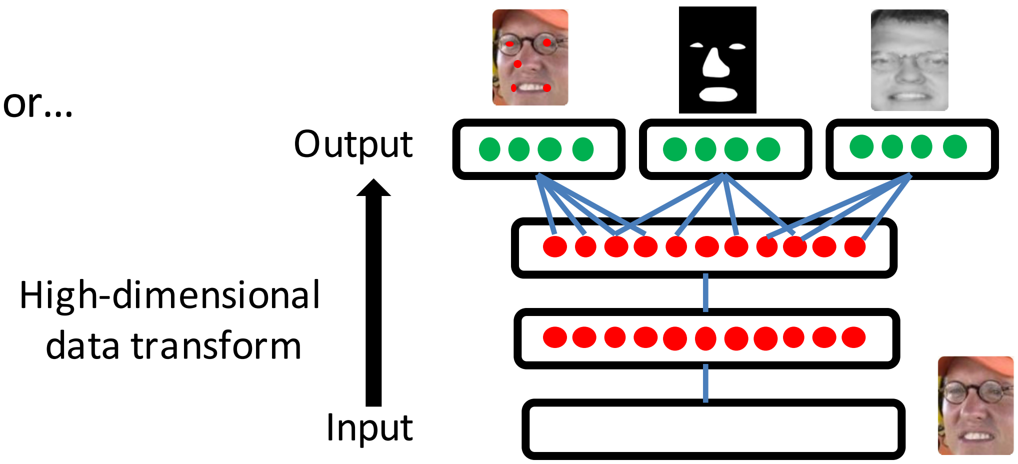
X. Zeng, W. Ouyang and X. Wang, " A Cascaded Deep Learning Architecture for Pedestrian Detection," ICCV 2013.

W. Ouyang and Xiaogang Wang, "Joint Deep Learning for Pedestrian Detection," IEEE ICCV 2013.

Large learning capacity makes high dimensional data transforms possible, and makes better use of contextual information

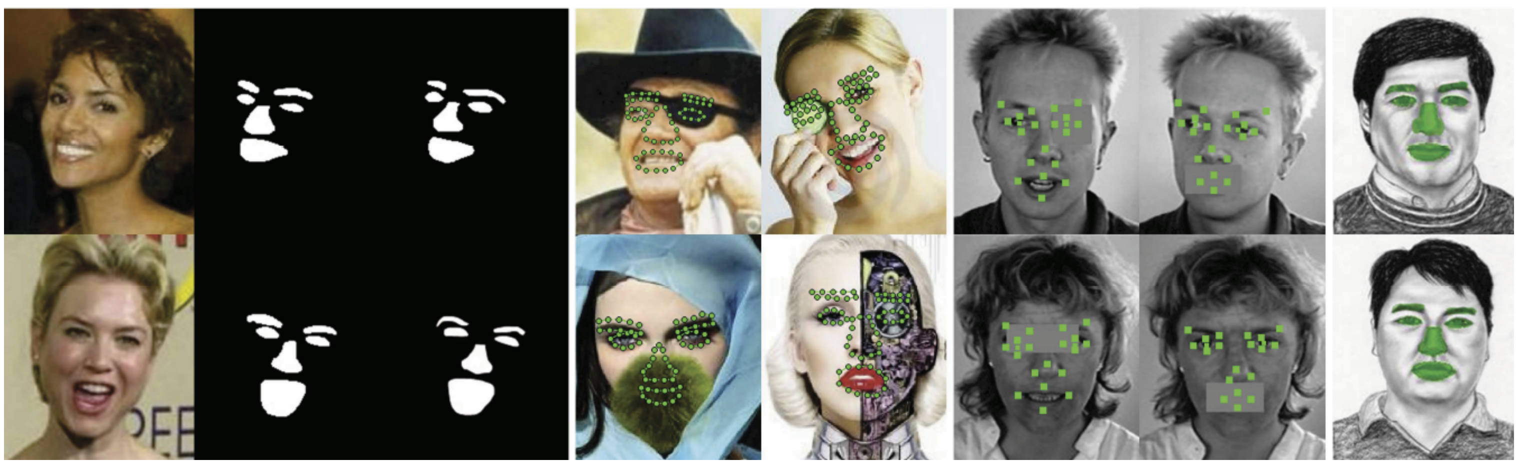
- How to make use of the large learning capacity of deep models?
  - **High dimensional data transform**
  - Hierarchical nonlinear representations



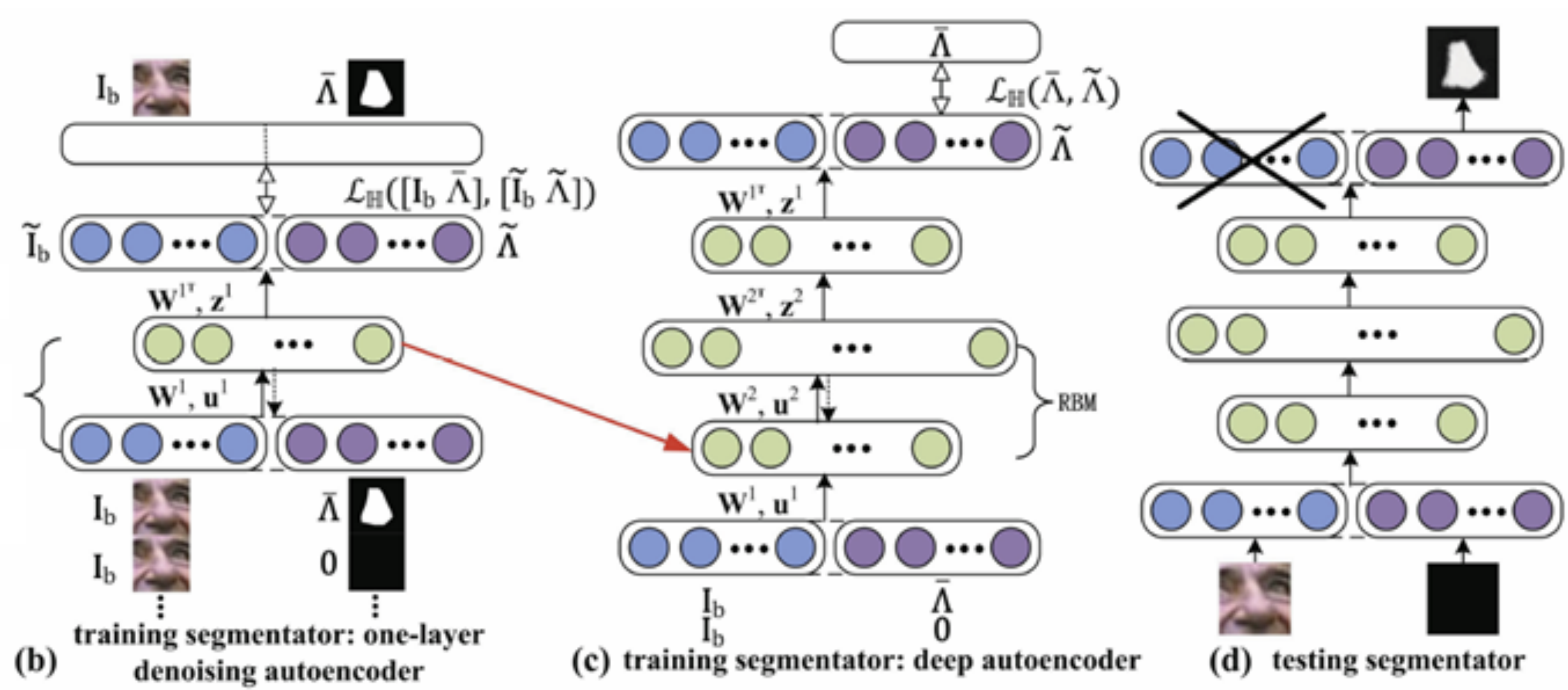


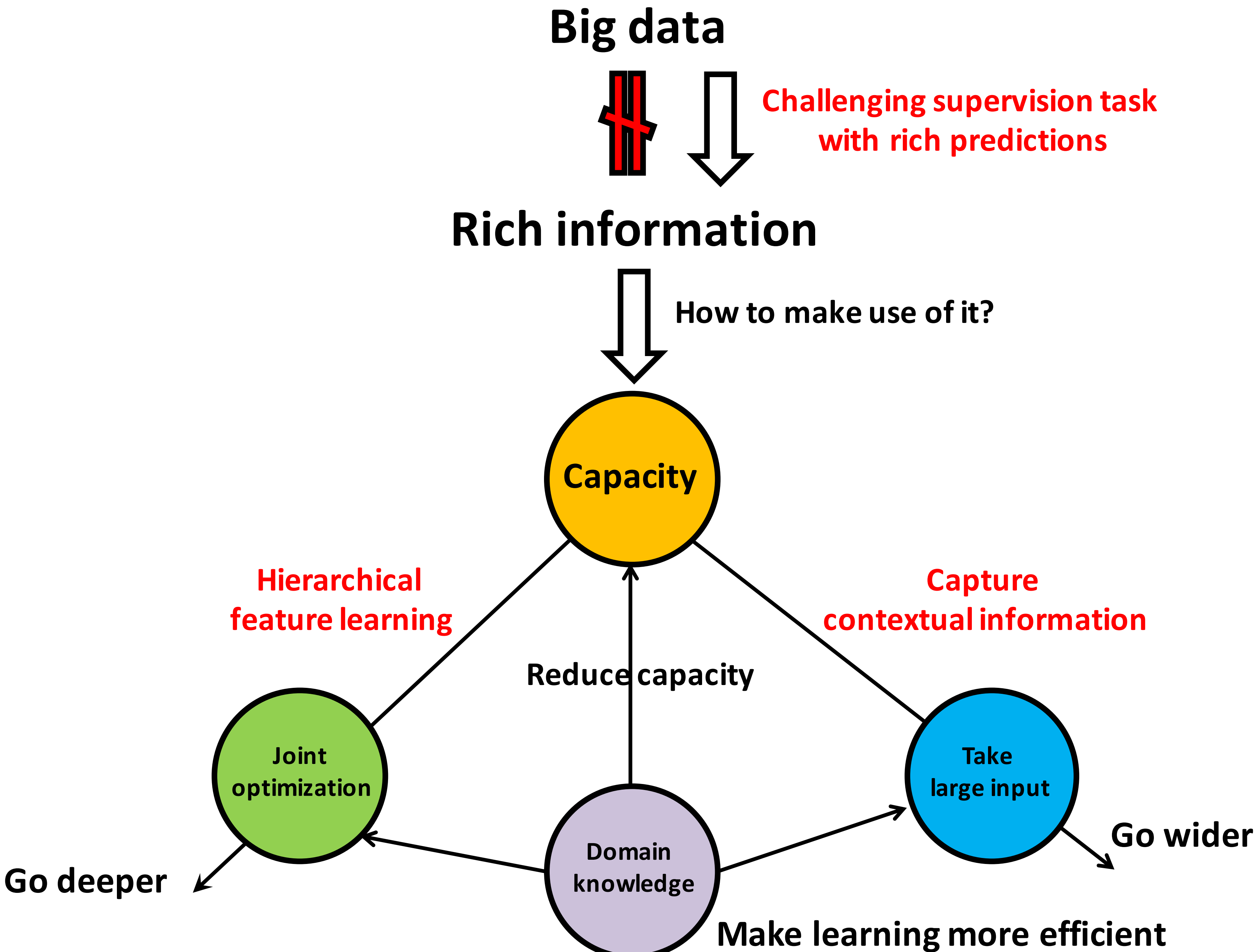
Face Parsing

- P. Luo, X. Wang and X. Tang, “Hierarchical Face Parsing via Deep Learning,” CVPR 2012



Training Segmentators



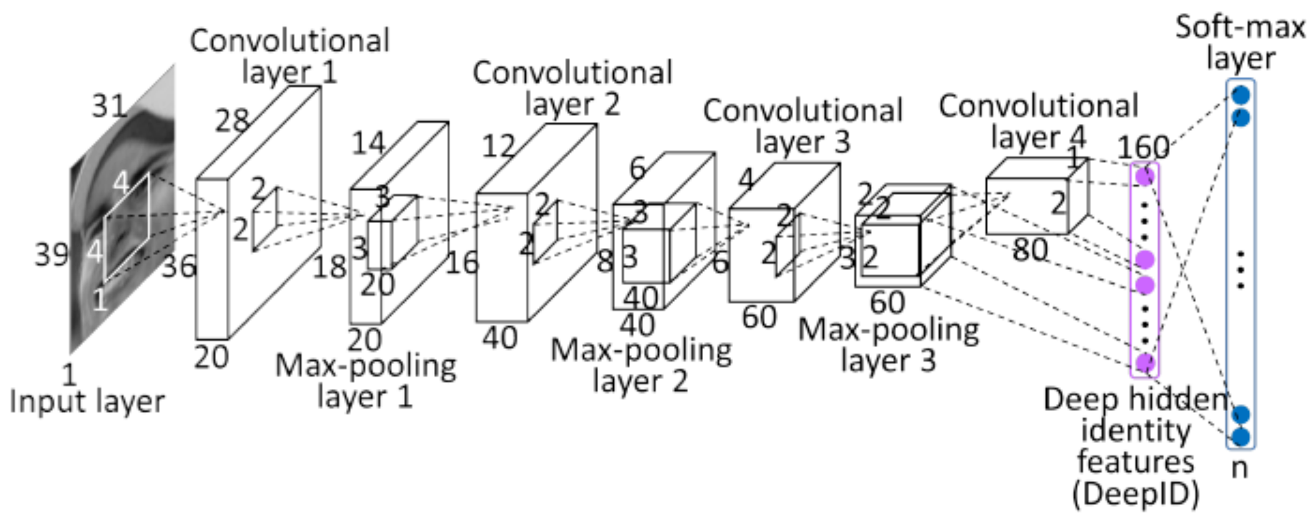

Big data
Challenging supervision task
with rich predictions
Rich information
How to make use of it?
Capacity
Hierarchical
feature learning
Capture
contextual information
Reduce capacity
Joint
optimization
Take
large input
Go wider
Go deeper
Domain
knowledge
Make learning more efficient

Outline

- Historical review of deep learning
- Understand deep learning
- **Interpret neural semantics**

DeepID2: Joint Identification (Id)-Verification (Ve) Signals

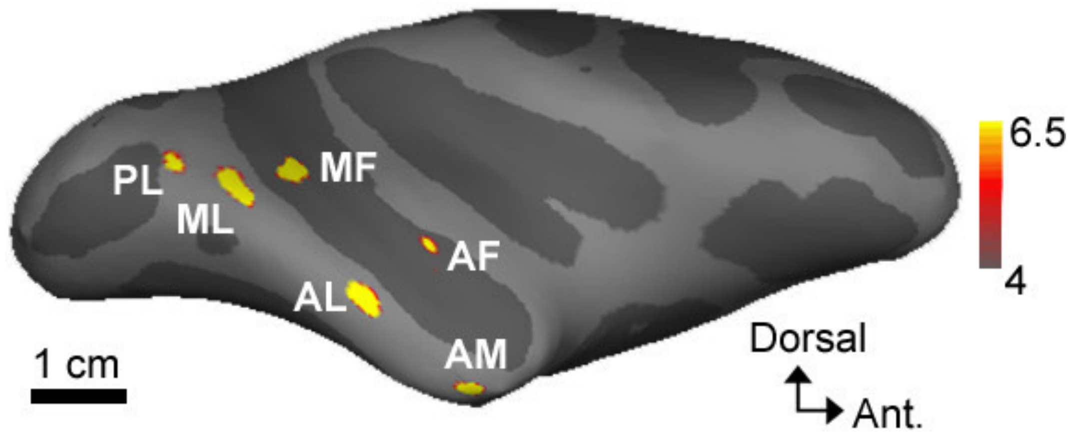
$$\mathrm{Verif}(f_i, f_j, y_{ij}, \theta_{ve}) = \begin{cases} \frac{1}{2}\left\|f_i - f_j\right\|_2^2 & \text{if } y_{ij} = 1 \\ \frac{1}{2}\max\left(0, m - \left\|f_i - f_j\right\|_2\right)^2 & \text{if } y_{ij} = -1 \end{cases}$$



(Id)

Y. Sun, X. Wang, and X. Tang. NIPS, 2014.

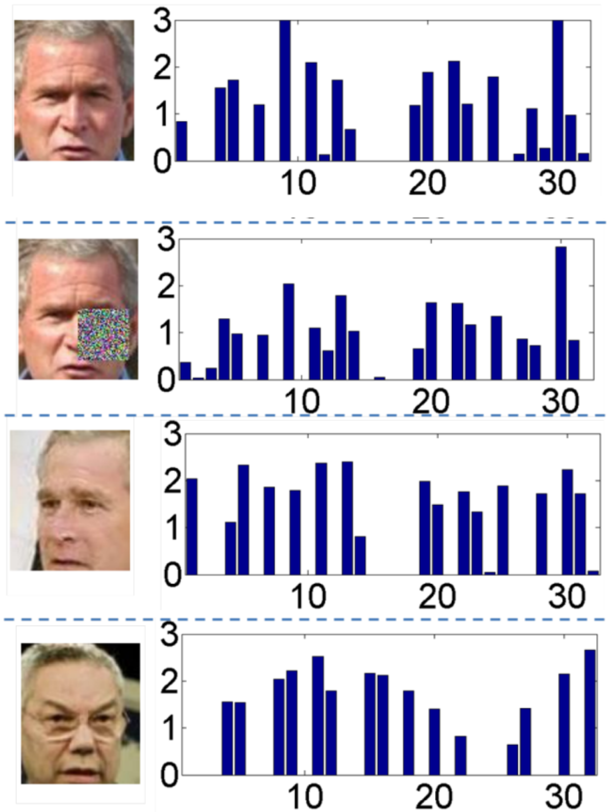
Biological Motivation



- Monkey has a face-processing network that is made of six interconnected face-selective regions
- Neurons in some of these regions were view-specific, while some others were tuned to identity across views
- View could be generalized to other factors, e.g. expressions?

Winrich A. Freiwald and Doris Y. Tsao, “Functional compartmentalization and viewpoint generalization within the macaque face-processing system,” *Science,* 330(6005):845–851, 2010.

Deeply learned features are moderately sparse

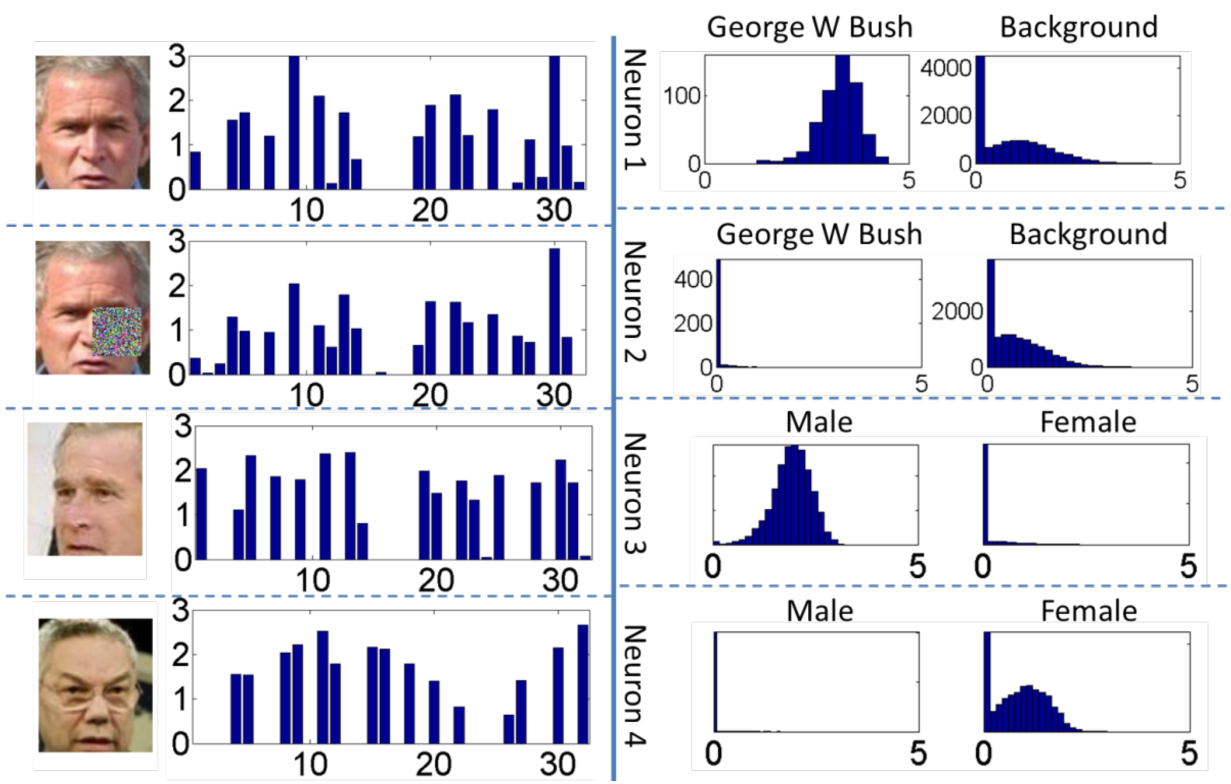


- The **binary codes** on activation patterns are very effective on face recognition
- Save storage and speedup face search dramatically
- Activation patterns are more important than activation magnitudes in face recognition

	Joint Bayesian (%)	Hamming distance (%)
Combined model (real values)	99.47	n/a
Combined model (binary code)	99.12	97.47

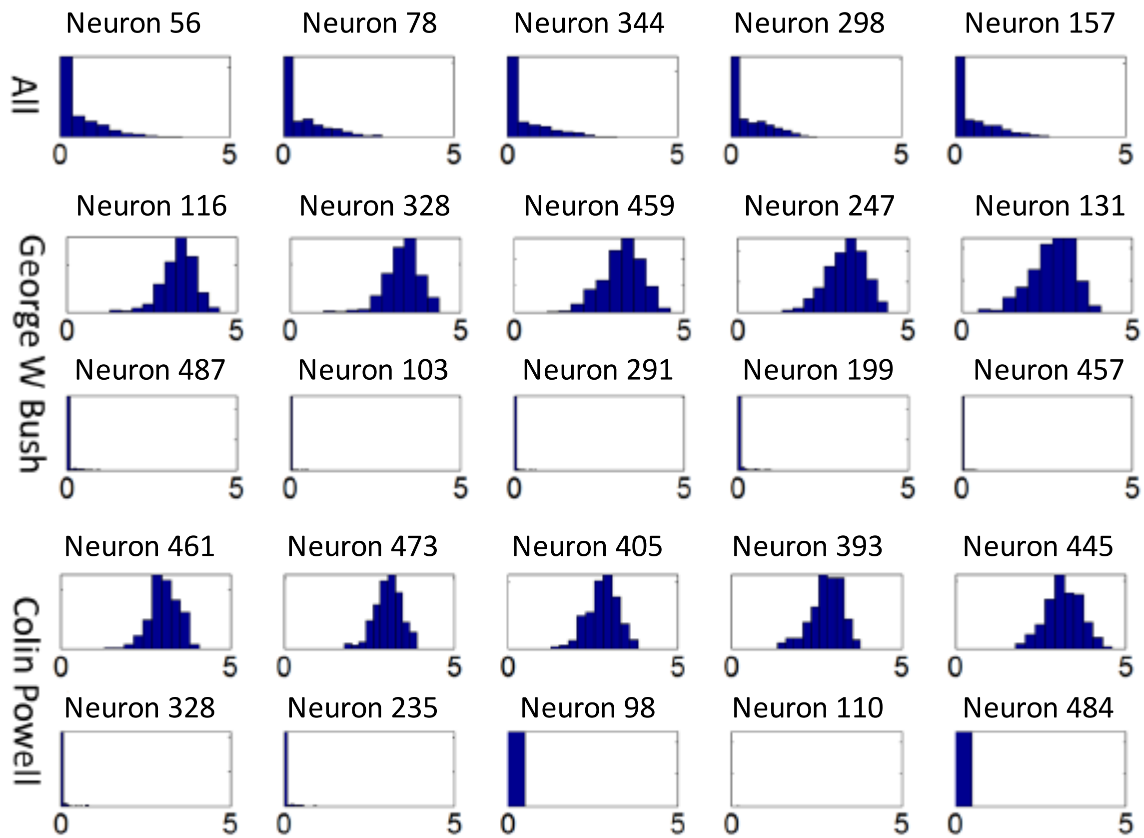
Deeply learned features are selective to identities and attributes

- With a single neuron, DeepID2 reaches 97% recognition accuracy for some identity and attribute

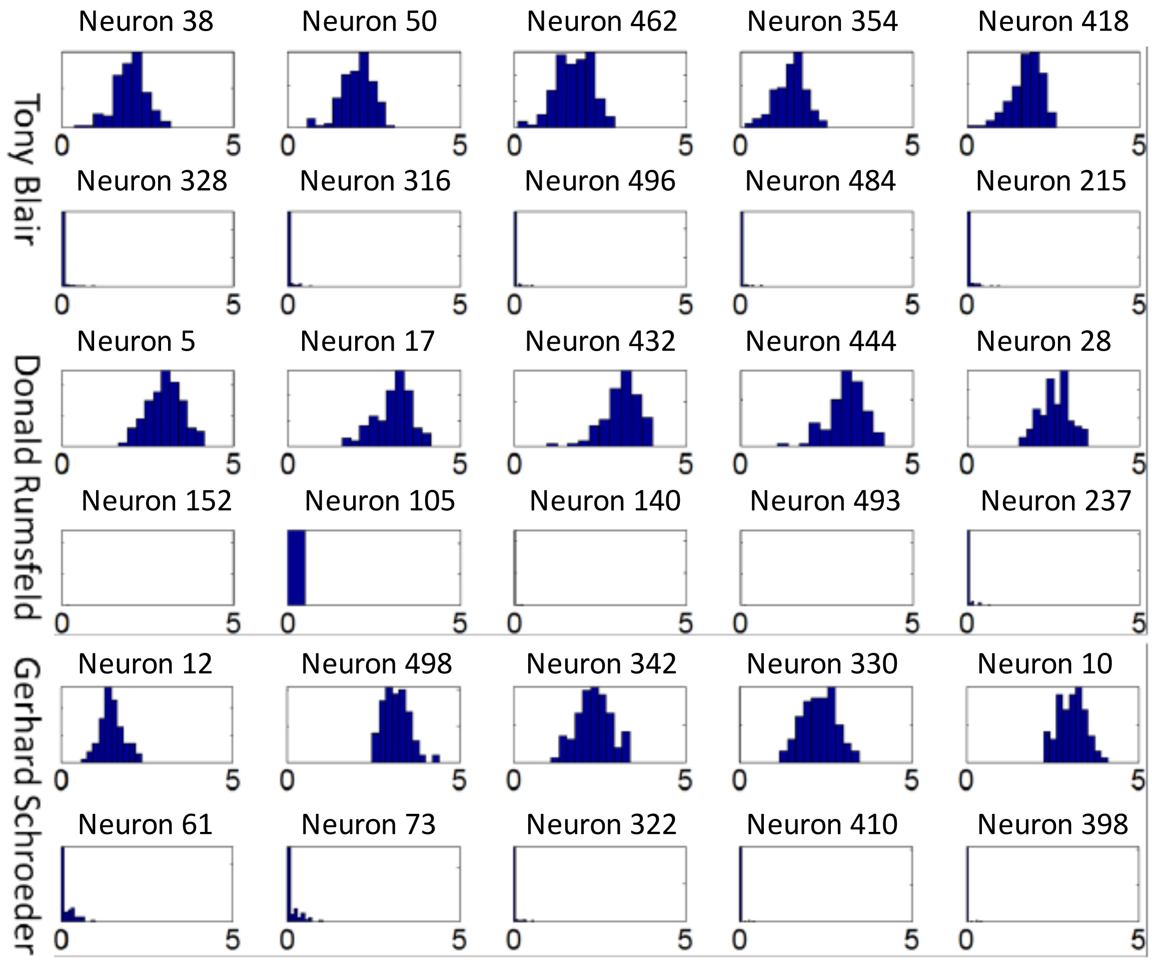


Deeply learned features are selective to identities and attributes

- Excitatory and inhibitory neurons (on identities)

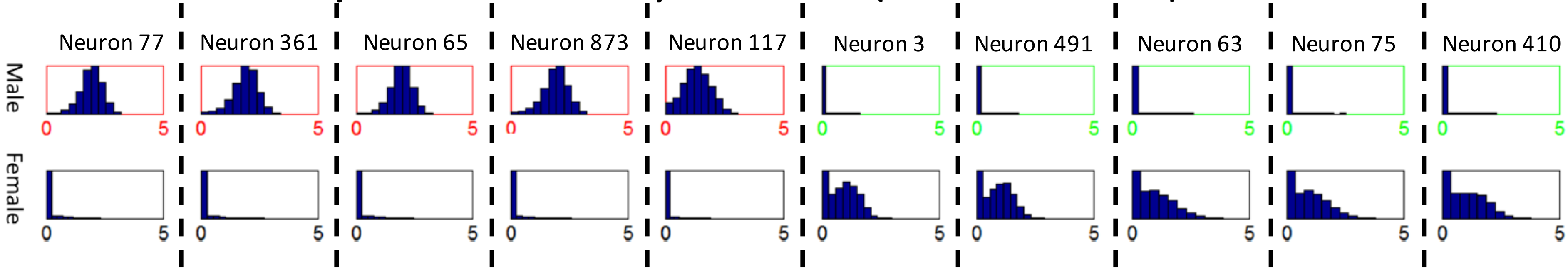


Histograms of neural activations over identities with the most images in LFW

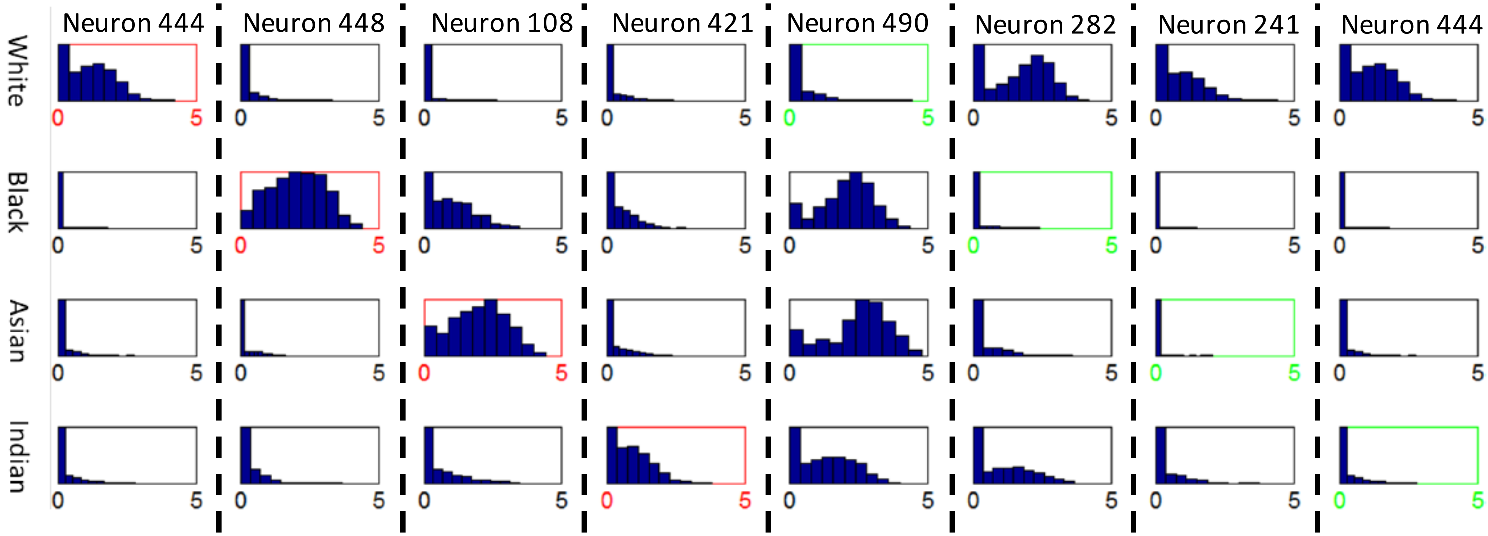

Tony Blair
Donald Rumsfeld
Gerhard Schroeder
Neuron 38
Neuron 50
Neuron 462
Neuron 354
Neuron 418
Neuron 328
Neuron 316
Neuron 496
Neuron 484
Neuron 215
Neuron 5
Neuron 17
Neuron 432
Neuron 444
Neuron 28
Neuron 152
Neuron 105
Neuron 140
Neuron 493
Neuron 237
Neuron 12
Neuron 498
Neuron 342
Neuron 330
Neuron 10
Neuron 61
Neuron 73
Neuron 322
Neuron 410
Neuron 398
0
5

Deeply learned features are selective to identities and attributes

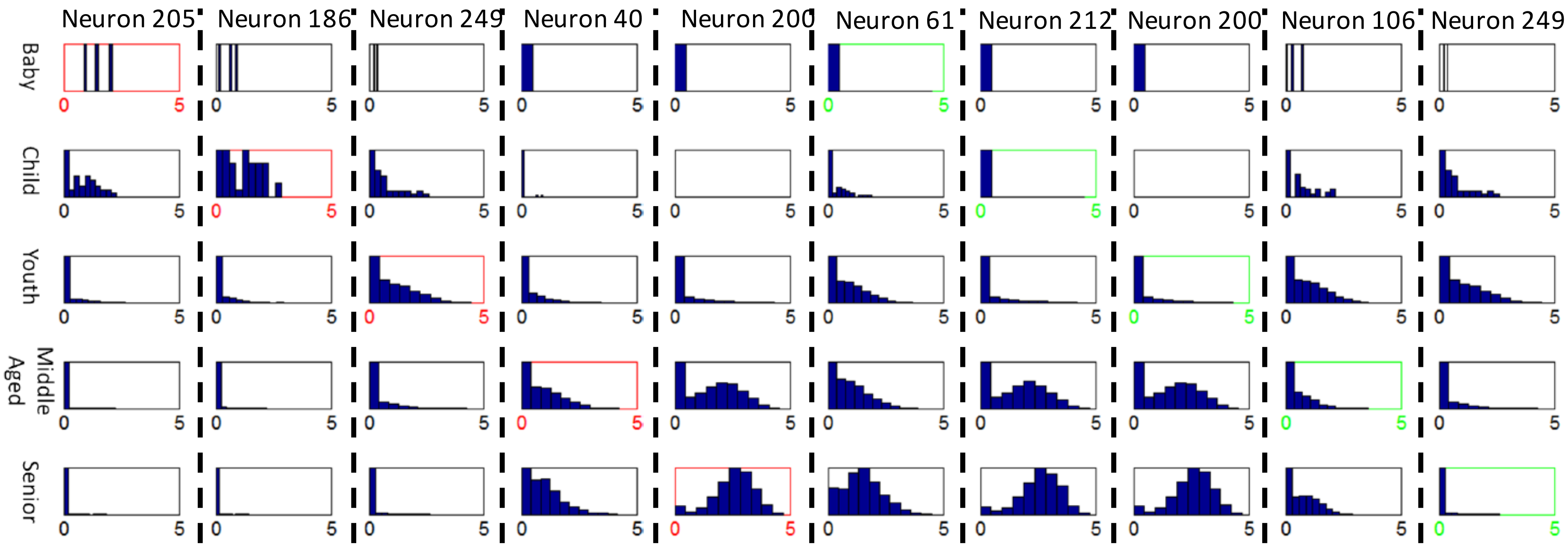
- Excitatory and inhibitory neurons (on attributes)



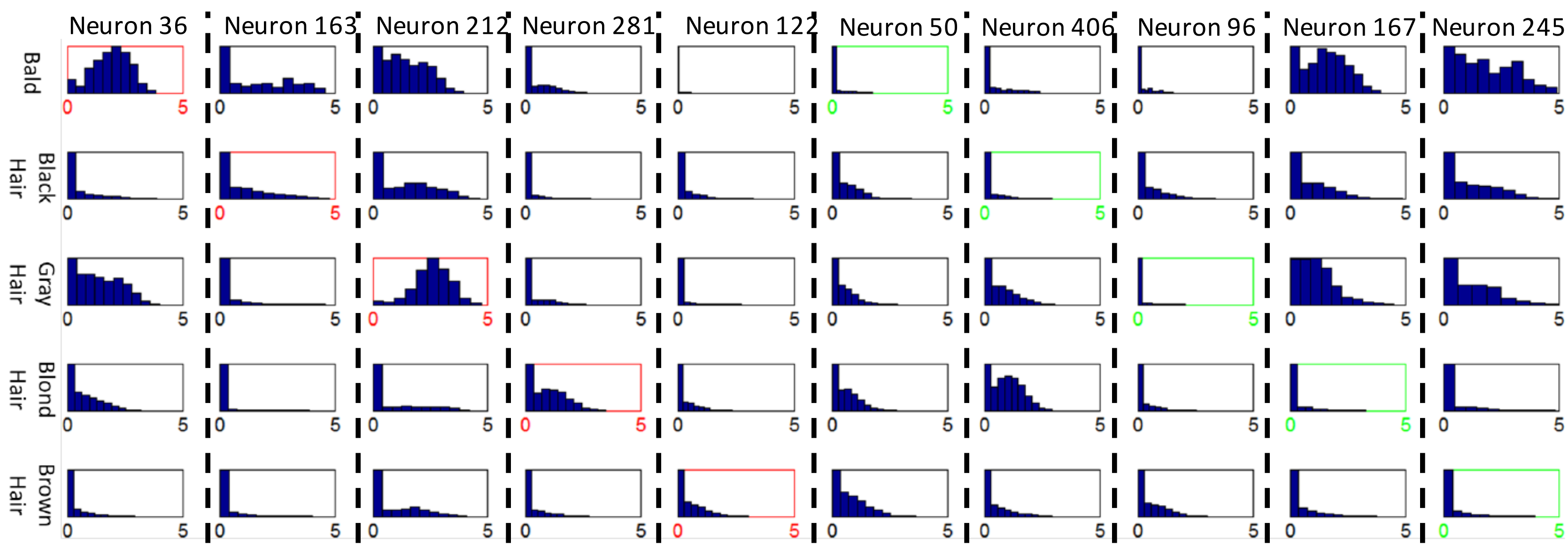
Histograms of neural activations over gender-related attributes (Male and Female)



Histograms of neural activations over race-related attributes (White, Black, Asian and India)



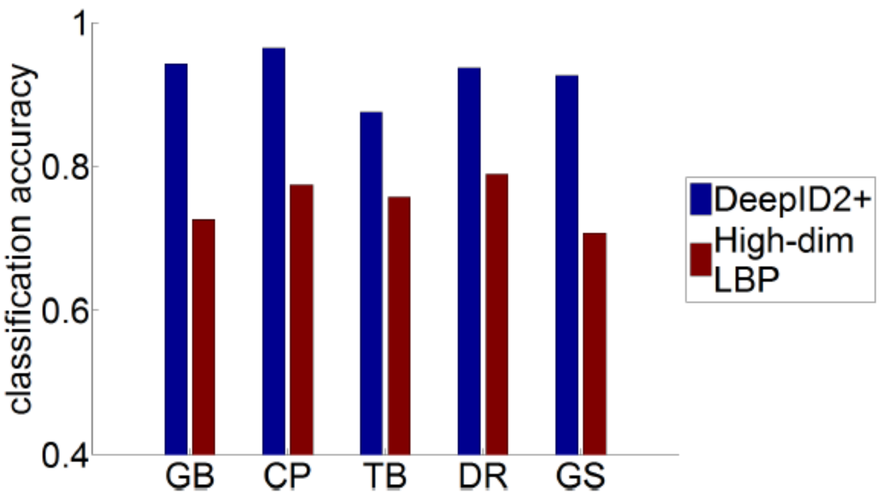
Histogram of neural activations over age-related attributes (Baby, Child, Youth, Middle Aged, and Senior)



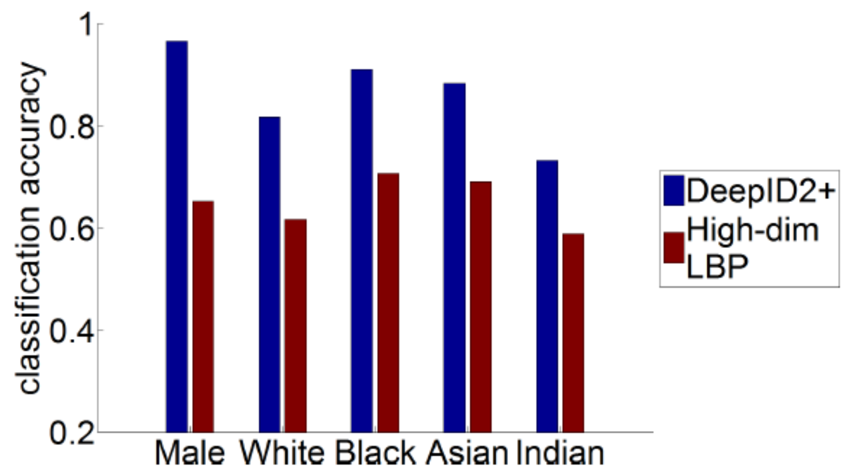
Histogram of neural activations over hair-related attributes (Bald, Black Hair, Gray Hair, Blond Hair, and Brown Hair.

Deeply learned features are selective to identities and attributes

- With a single neuron, DeepID2 reaches 97% recognition accuracy for some identity and attribute



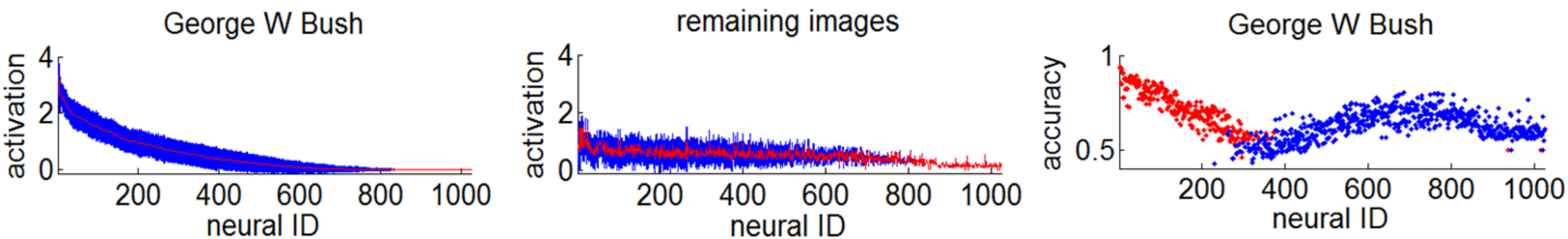
Identity classification accuracy on LFW with one single DeepID2+ or LBP feature. GB, CP, TB, DR, and GS are five celebrities with the most images in LFW.



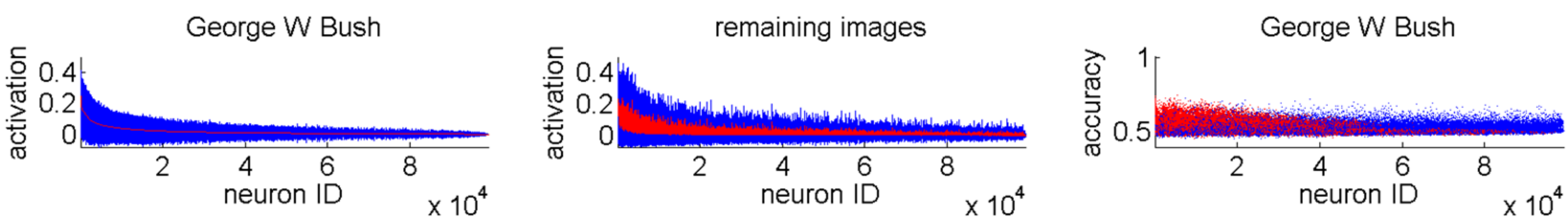
Attribute classification accuracy on LFW with one single DeepID2+ or LBP feature.

Excitatory and Inhibitory neurons

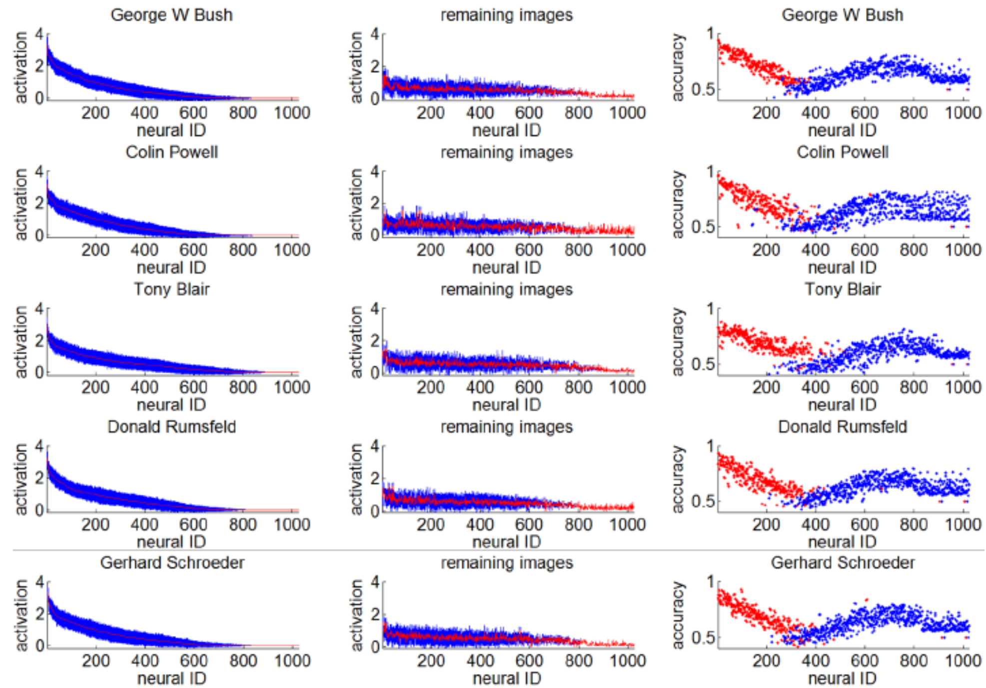
DeepID2+



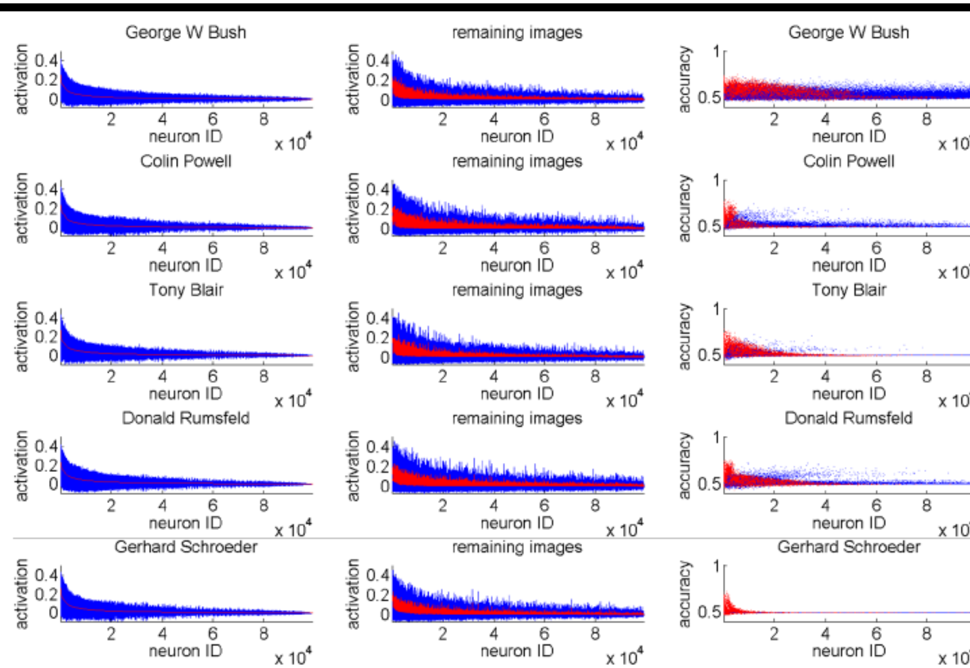
High-dim LBP



Excitatory and Inhibitory neurons

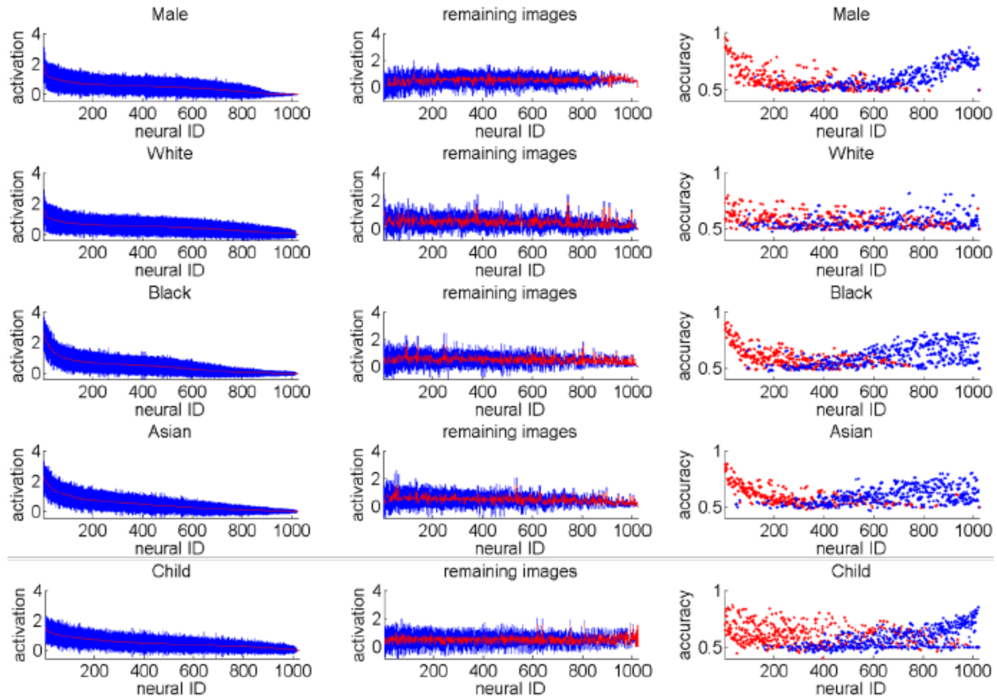


DeepID2+

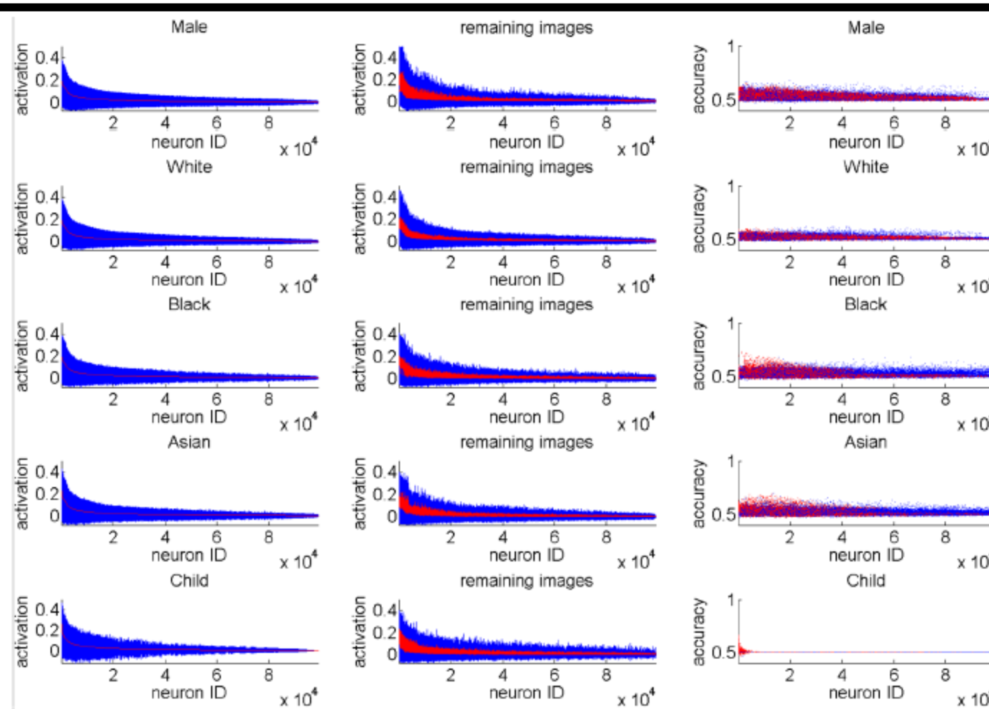


High-dim LBP

Excitatory and Inhibitory neurons



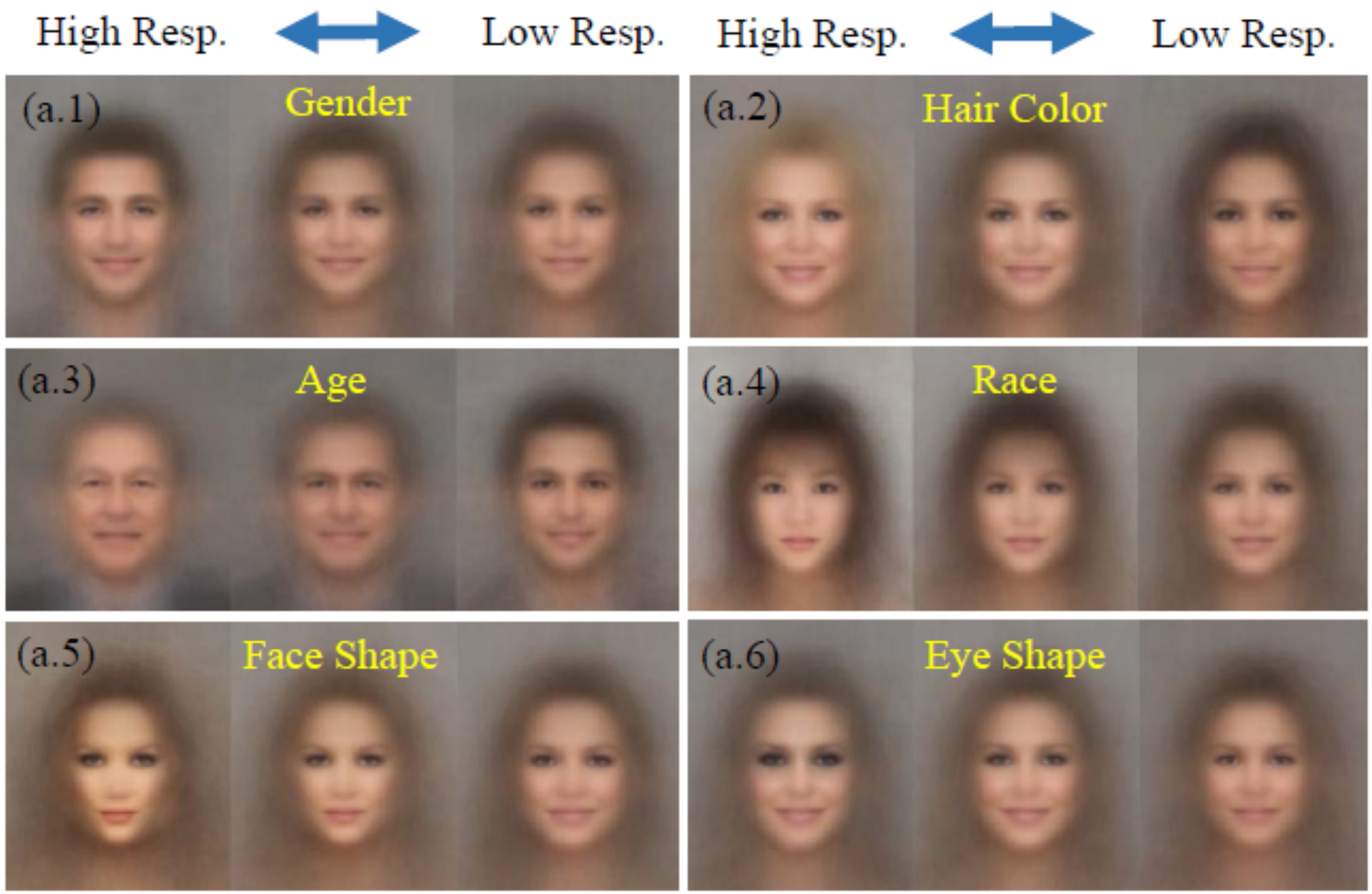
DeepID2+

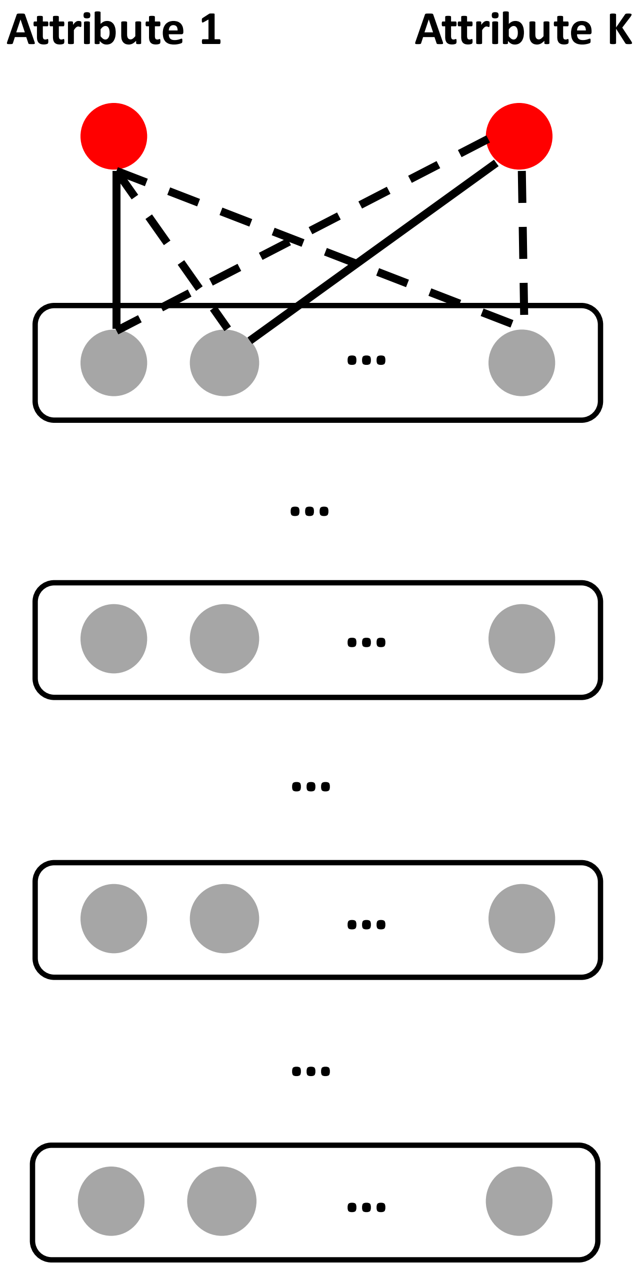


High-dim LBP

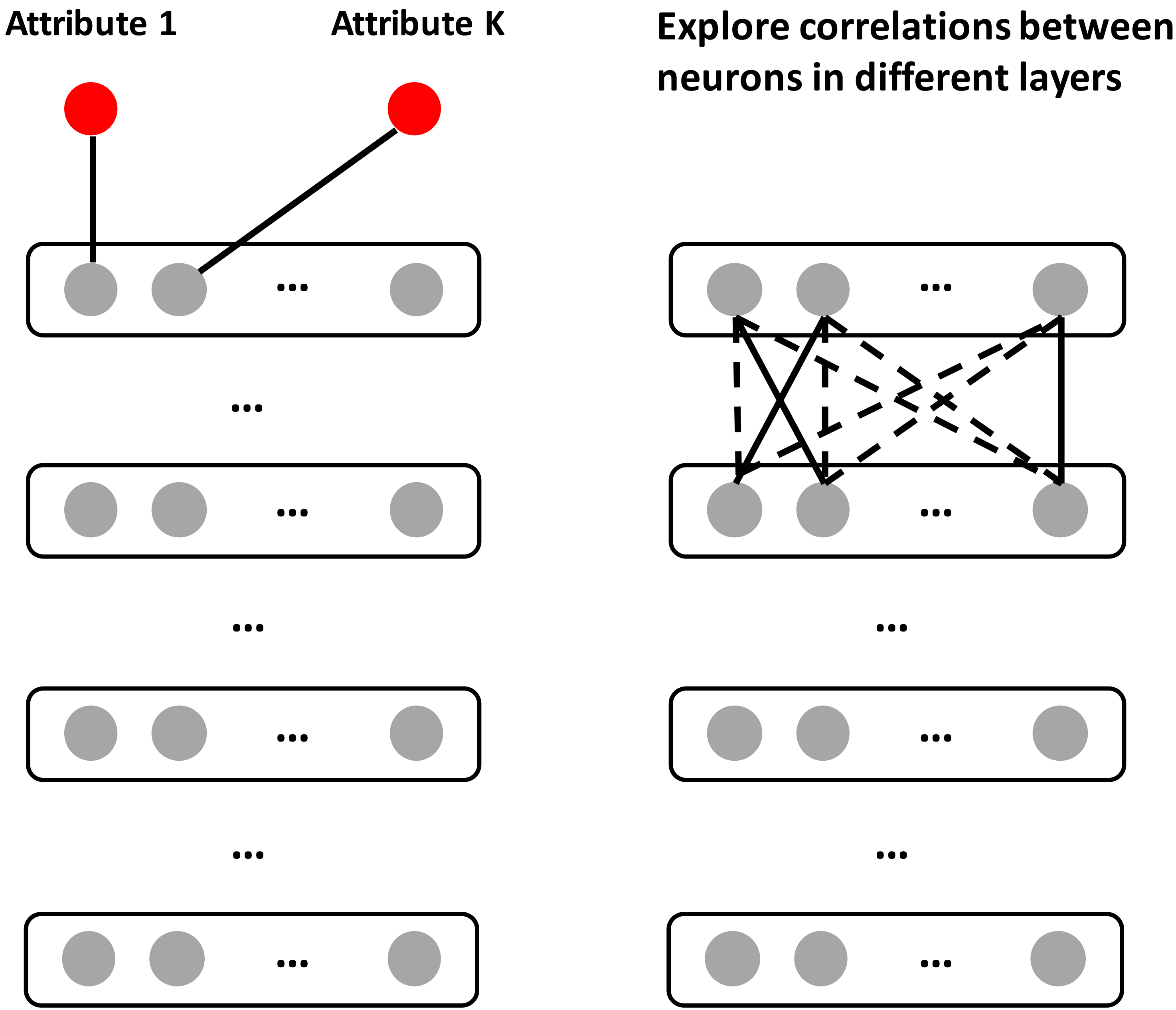
Deeply learned features are selective to identities and attributes

- Visualize the semantic meaning of each neuron

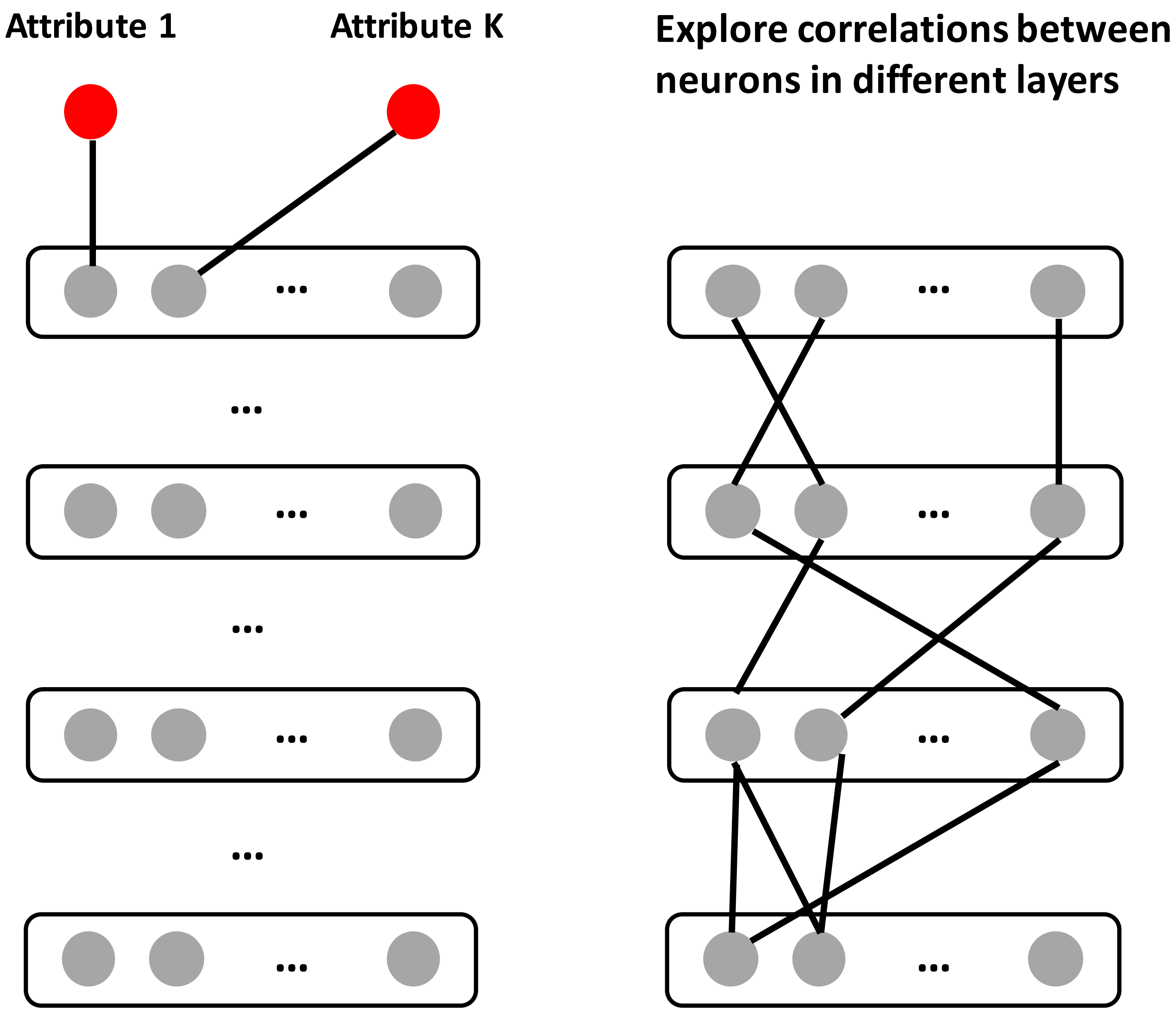




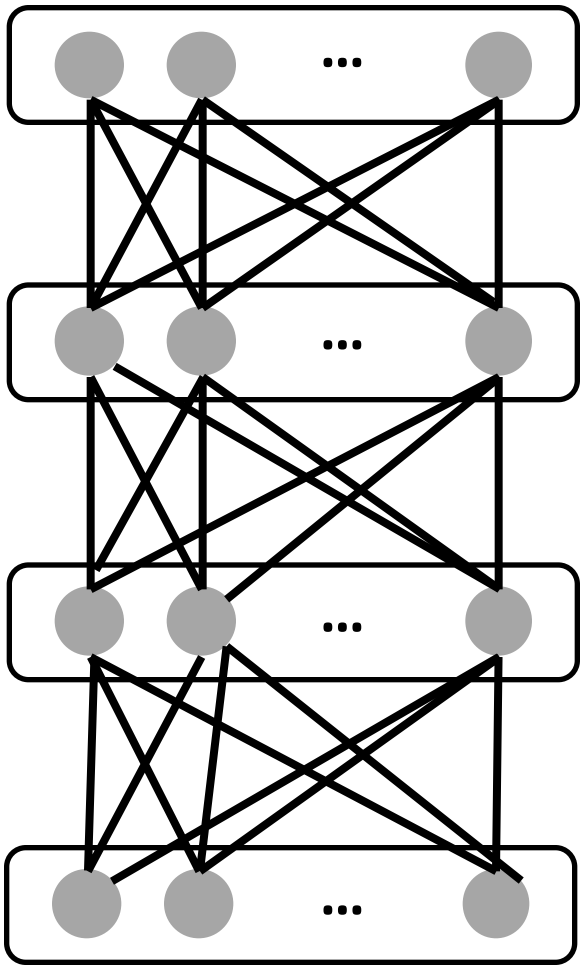
Yi Sun, Xiaogang Wang, and Xiaoou Tang, “Sparsifying Neural Network Connections for Face Recognition,” arXiv:1512.01891, 2015



Attribute 1
Attribute K
Explore correlations between neurons in different layers


Attribute 1
Attribute K
Explore correlations between neurons in different layers

Alternatively learning weights and net structures



1. Train a dense network from scratch
2. Sparsify the top layer, and **re-train** the net
3. Sparsify the second top layer, and **re-train** the net

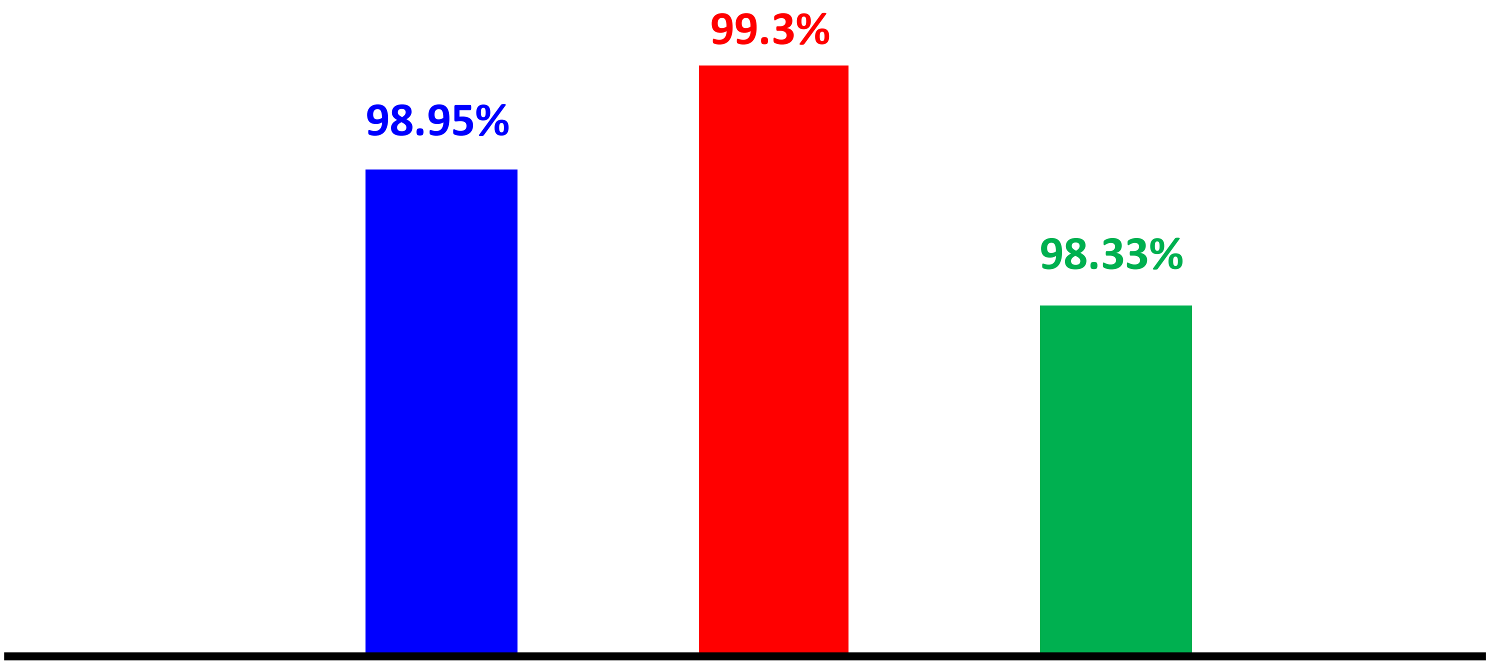
…

Conel, JL. The postnatal development of the human cerebral cortex. Cambridge, Mass: Harvard University Press, 1959.

Original deep neural network

Sparsified deep neural network and only keep 1/8 amount of parameters after joint optimization of weights and structures

Train the sparsified network from scratch



The sparsified network has enough learning capacity, but the original denser network helps it reach a better intialization

Deep learning = ?

Machine learning with big data

Feature learning

Joint learning

Contextual learning

Deep feature presentations are

Sparse

Selective

Robust to data corruption

References

- D. E. Rumelhart, G. E. Hinton, R. J. Williams, “Learning Representations by Back-propagation Errors,” Nature, Vol. 323, pp. 533-536, 1986.
- N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. Piater, A. J. Rodriguez-Sanchez, L. Wiskott, “Deep Hierarchies in the Primate Visual Cortex: What Can We Learn For Computer Vision?” IEEE Trans. PAMI, Vol. 35, pp. 1847-1871, 2013.
- A. Krizhevsky, L. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” Proc. NIPS, 2012.
- Y. Sun, X. Wang, and X. Tang, “Deep Learning Face Representation by Joint Identification-Verification,” NIPS, 2014.
- K. Fukushima, “Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position,” Biological Cybernetics, Vol. 36, pp. 193-202, 1980.
- Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based Learning Applied to Document Recognition,” Proceedings of the IEEE, Vol. 86, pp. 2278-2324, 1998.
- G. E. Hinton, S. Osindero, and Y. Teh, “A Fast Learning Algorithm for Deep Belief Nets,” Neural Computation, Vol. 18, pp. 1527-1544, 2006.

- G. E. Hinton and R. R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Networks,” Science, Vol. 313, pp. 504-507, July 2006.
- Z. Zhu, P. Luo, X. Wang, and X. Tang, “Deep Learning Identity Face Space,” Proc. ICCV, 2013.
- Z. Zhu, P. Luo, X. Wang, and X. Tang, “Deep Learning and Disentangling Face Representation by Multi-View Perception,” NIPS 2014.
- Y. Sun, X. Wang, and X. Tang, “Deep Learning Face Representation from Predicting 10,000 classes,” Proc. CVPR, 2014.
- J. Hastad, “Almost Optimal Lower Bounds for Small Depth Circuits,” Proc. ACM Symposium on Theory of Computing, 1986.
- J. Hastad and M. Goldmann, “On the Power of Small-Depth Threshold Circuits,” Computational Complexity, Vol. 1, pp. 113-129, 1991.
- A. Yao, “Separating the Polynomial-time Hierarchy by Oracles,” Proc. IEEE Symposium on Foundations of Computer Science, 1985.
- Sermnet, K. Kavukcuoglu, S. Chintala, and LeCun, “Pedestrian Detection with Unsupervised Multi-Stage Feature Learning,” CVPR 2013.
- W. Ouyang and X. Wang, “Joint Deep Learning for Pedestrian Detection,” Proc. ICCV, 2013.
- P. Luo, X. Wang and X. Tang, “Hierarchical Face Parsing via Deep Learning,” Proc. CVPR, 2012.
- Honglak Lee, “Tutorial on Deep Learning and Applications,” NIPS 2010.


Questions?