Convolutional Neural Networks for Computer Vision Applications

林彥宇 副研究員
Yen-Yu Lin, Associate Research Fellow
中央研究院 資訊科技創新研究中心
Research Center for IT Innovation, Academia Sinica

About Yen-Yu Lin

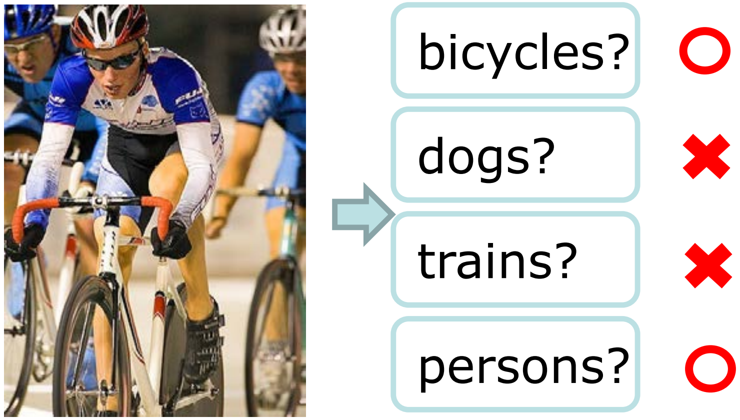
- Yen-Yu Lin, *Associate research fellow, CITI, Academia Sinica*

- Research interests:
  - Computer Vision (CV):

    Let computers see, recognize, and interpret the world like humans
  - Machine Learning (ML):

    A statistical way to learn how human visual system works
  - Goal: *Design **ML** methods to facilitate **CV** applications*

Research Topics 1/4

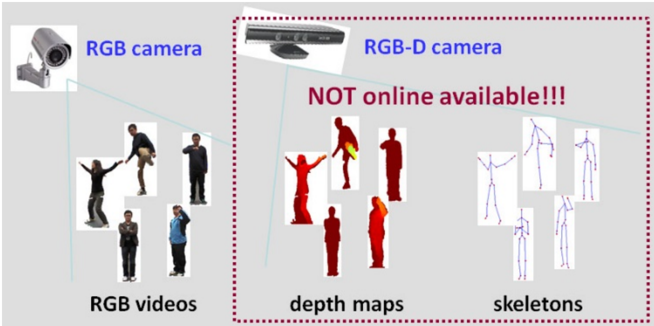


CV: *object recognition*
ML: *multiple kernel learning*
TPAMI'11, ICCV'09, NIPS'08

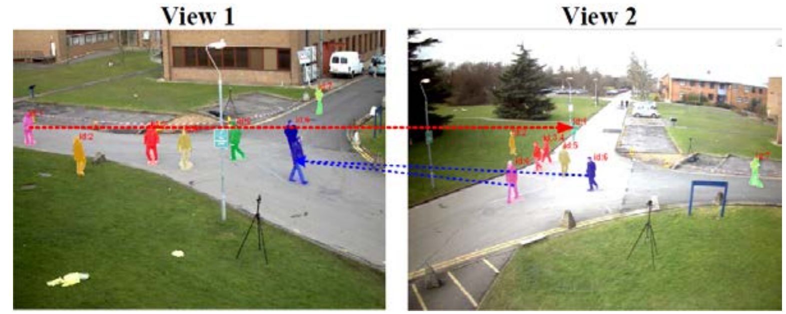
CV: *image segmentation*
ML: *graphical model*
CVPR'14, TIP'14, ACCV'12

CV: *face detection* **ML**: *multi-task boosting*
US Patent'07, CVPR'05, ECCV'04

Research Topics 2/4



CV: *action recognition*
ML: *low-rank reconstruction*
TIP'15, CVPR'14



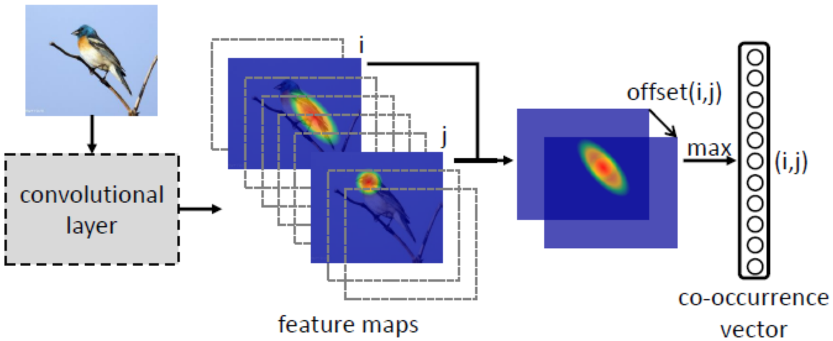
CV: *multi-view people counting*
ML: *transfer learning*
TIP'15, ACM MM'12

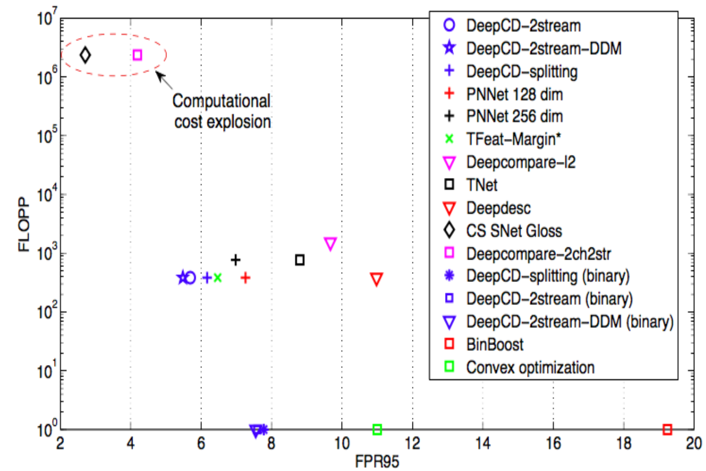
SIFT LIOP DAISY RI GB OURS!

CV: *image matching* **ML**: *energy minimization*

CVPR'16, TPAMI'15, TIP'15, CVPR'15, CVPR'13

Research Topics 3/4





CV: *fine-grained object recognition*

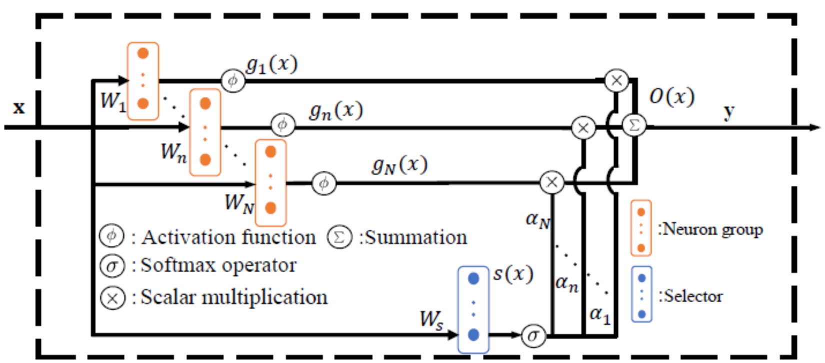
ML: *CNNs with co-occurrence layer*

CVPR'17

CV: *patch descriptor learning*

ML: *CNNs with adaptive learning rate*

ICCV'17

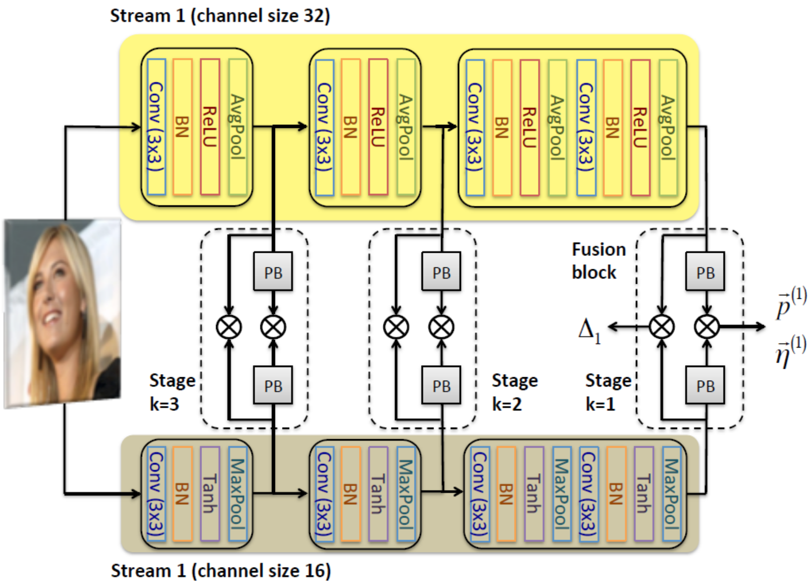


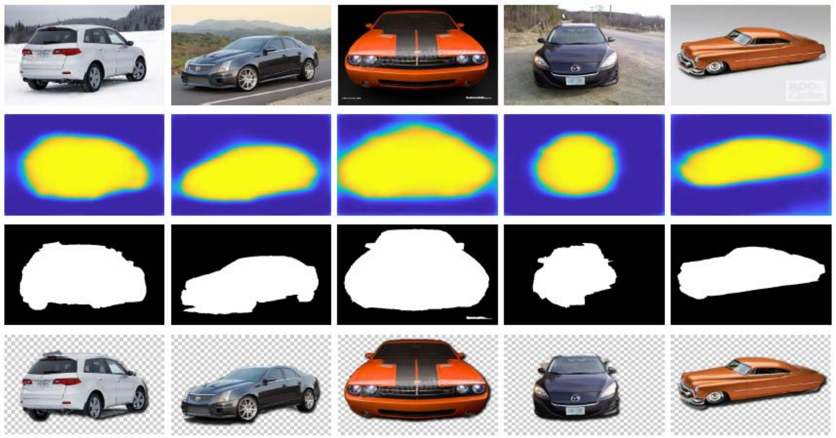
CV: *gesture recognition*

ML: *DNNs with adaptive hidden layer*

AAAI'18

Research Topics 4/4





CV: *face age estimation*
ML: *CNNs for hierarchical regression*
IJCAI'18

CV: *image co-segmentation*
ML: *Unsupervised CNNs*
IJCAI'18

Outline

- Convolutional neural networks (CNNs)
- Representative CNN models
- CNN-based computer vision applications

Outline

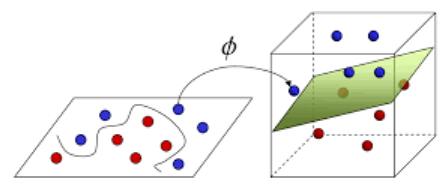
- Convolutional neural networks (CNNs)
  - Conventional approaches vs. deep learning
  - Neural networks
  - Convolutional neural networks
- Representative CNN models
- CNN-based computer vision applications

Conventional approach to object recognition

- Training phase

image collection → feature extraction → classifier training → trained classifier

histogram of oriented gradients (HoG)



support vector machines (SVMs)

dogs?	✖
flowers?	○
trains?	✖
persons?	✖

Conventional approach to object recognition

key step

- Training phase

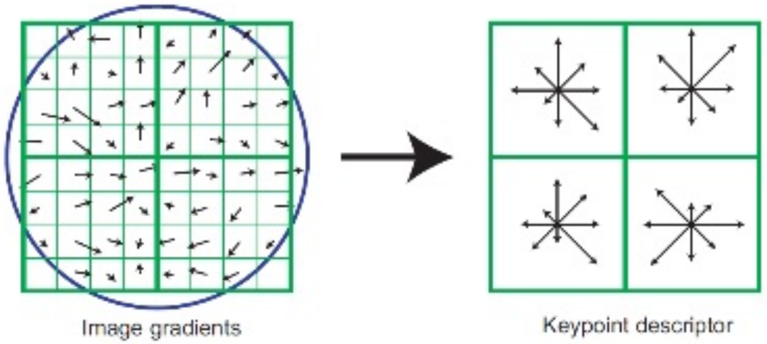
image collection → feature extraction → classifier training → trained classifier

- Testing phase

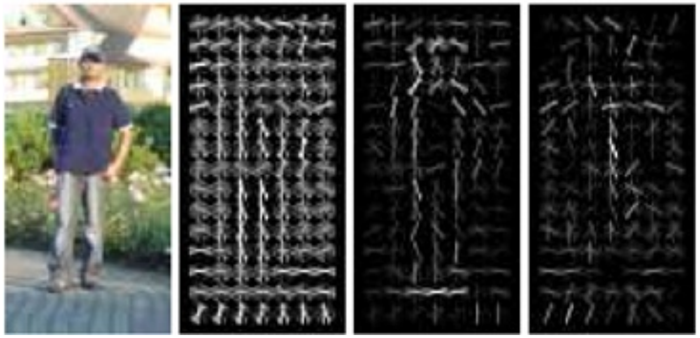
test image → feature extraction → trained classifier → prediction

Features are the keys

- Off-the-shelf visual features

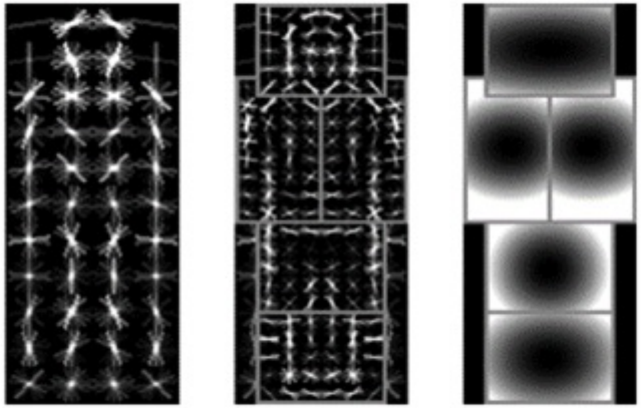


SIFT [Lowe, IJCV'04]
Citations: 43465



HoG [Dalal & Triggs, CVPR'05]
Citations: 20174

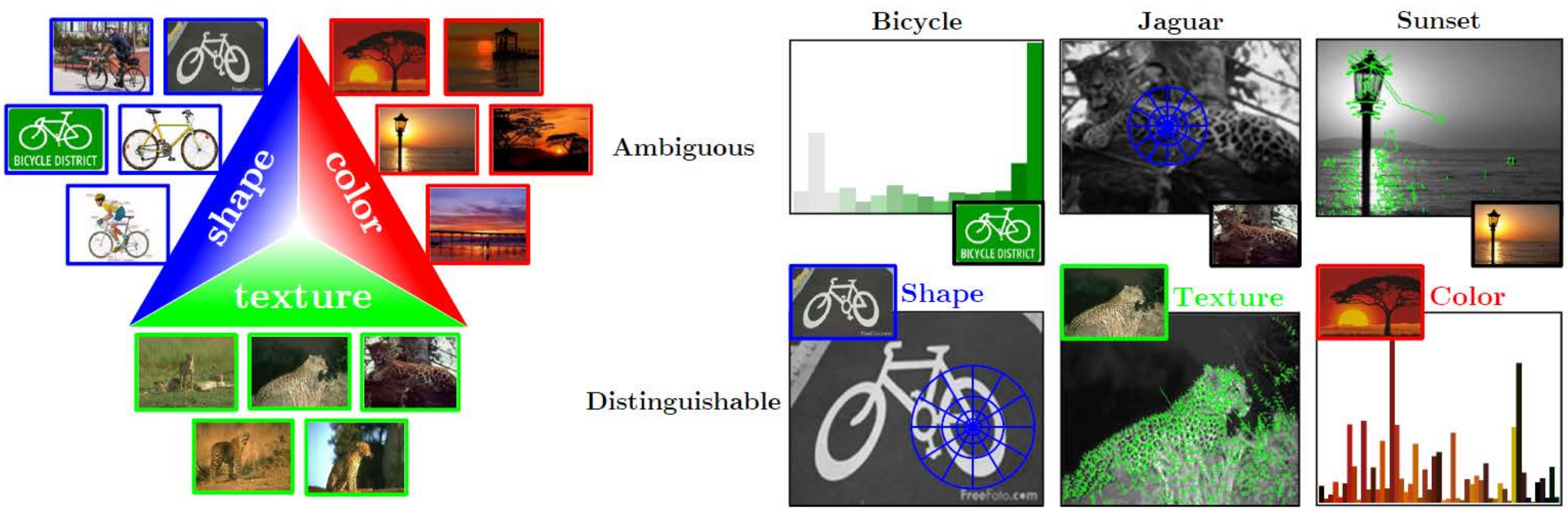
Constellation model [Fergus et al., CVPR'03]
Citations: 2551



DPM [Felzenszwalb et al., PAMI'10]
Citations: 5093

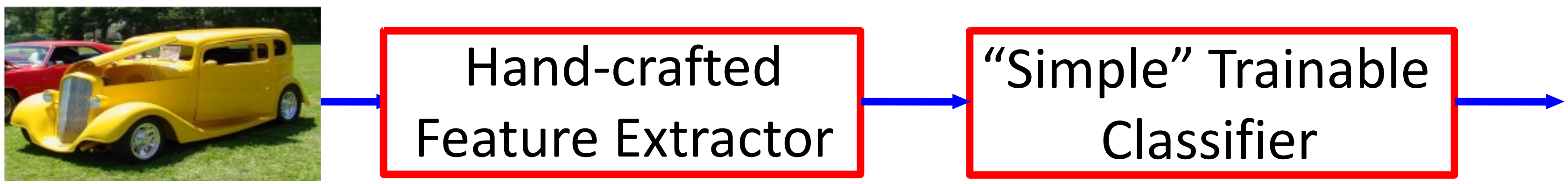
Features are the keys

- Features are the keys to recent progress in classification
- Are handcrafted features optimal?
- The optimal features for classification in general vary from task to task, even from category to category

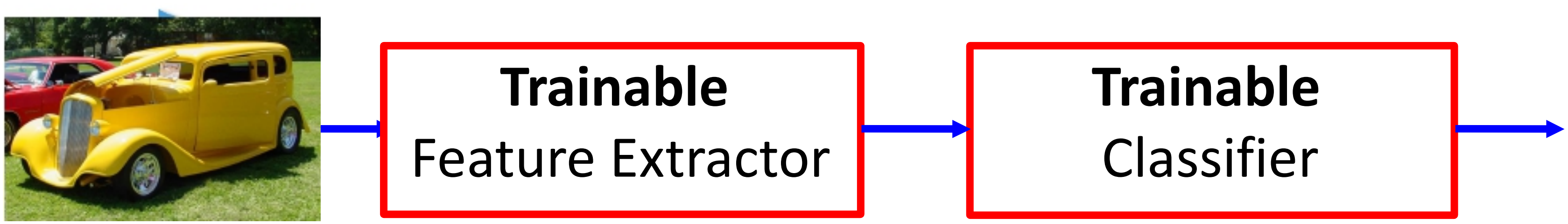


Conventional approaches vs. Deep learning

- Conventional approaches
  - Fixed/engineered features + trainable classifier

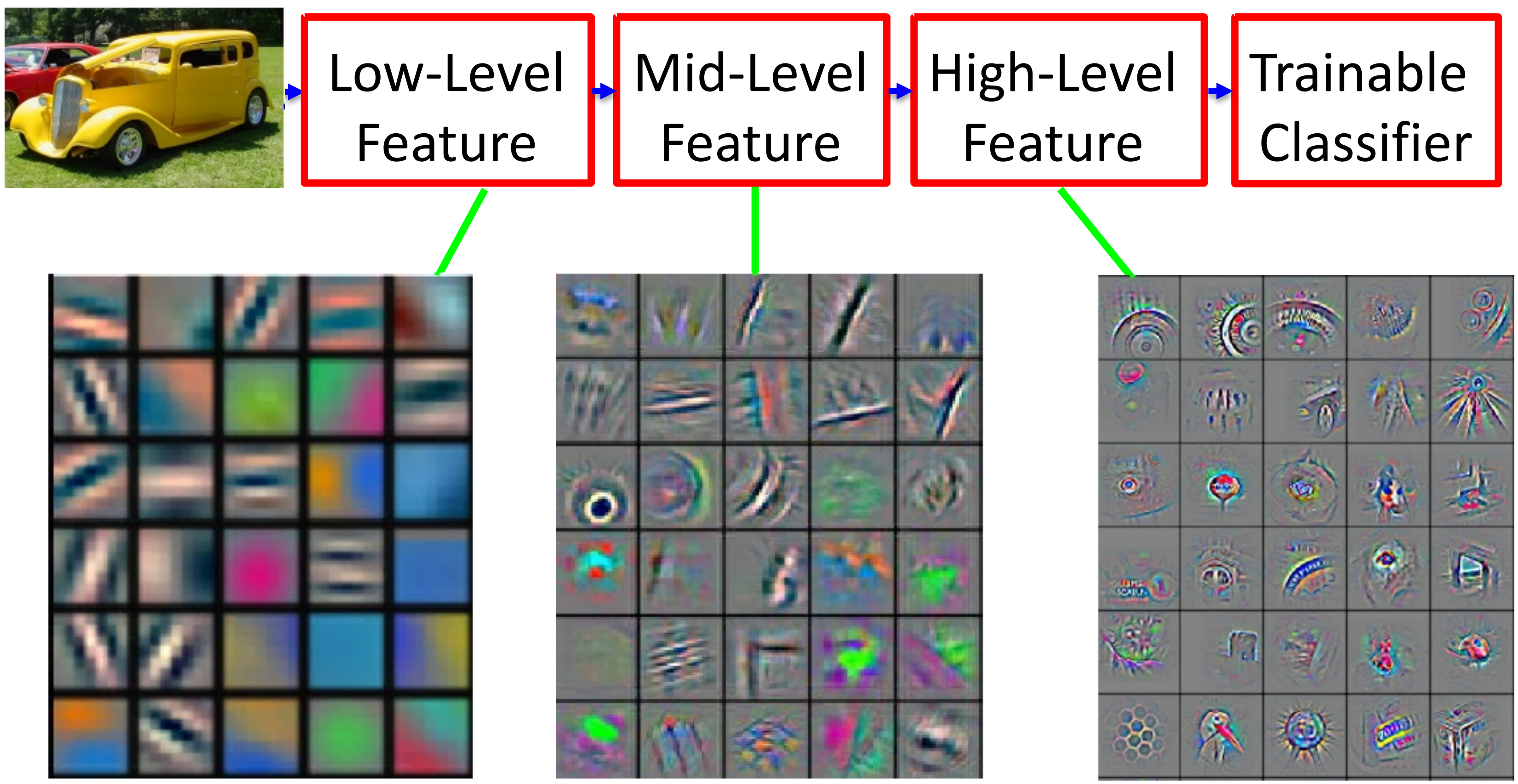


- Deep learning / End-to-end learning / Feature learning
  - Trainable features + trainable classifier



slide: Y LeCun & MA Ranzato

Deep learning = Learning hierarchical representations



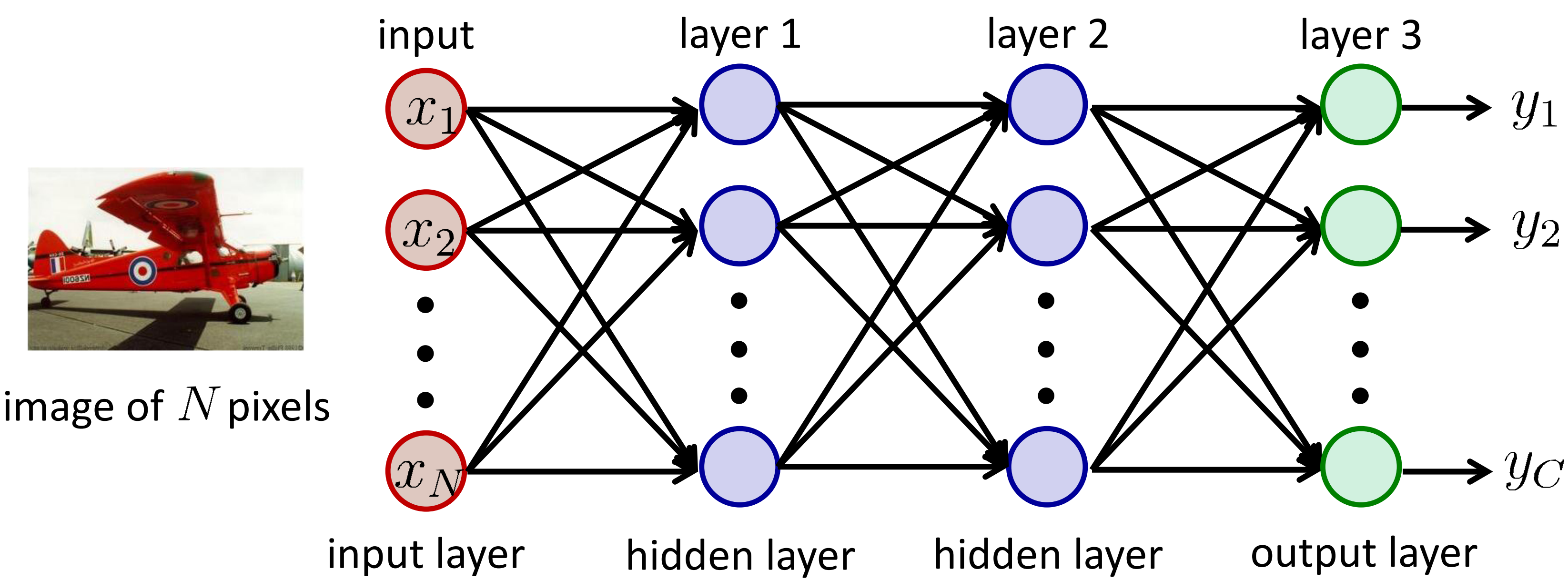
slide: Y LeCun & MA Ranzato

Outline

- Convolutional neural networks (CNN)
  - Conventional approaches vs. deep learning
  - Neural networks
  - Convolutional neural networks
- Representative CNN models
- CNN-based computer vision applications

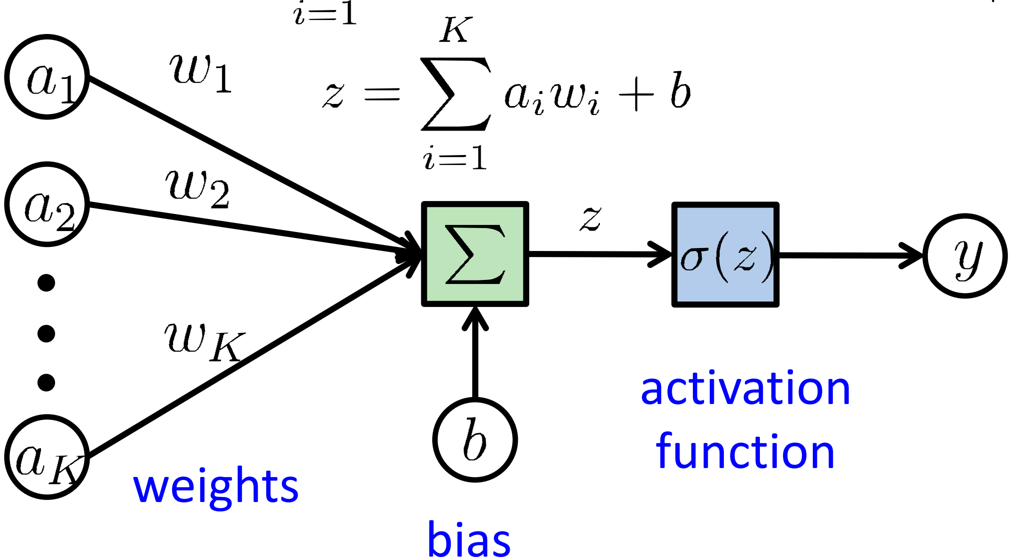
Neural networks and neurons

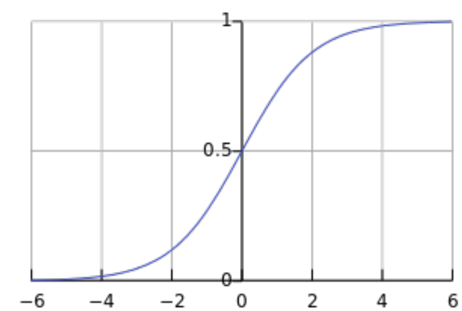
- Neural networks are presented as layers of interconnected neurons
  - Each layer of neurons takes messages from output of previous layer



A single neuron

- A function $f : R^K \mapsto R$
  - Map K inputs to 1 output
  - Compute the biased weighted sum
  - Apply a non-linear mapping function (activation function)
  - $f((a)) = \sigma(\sum_{i=1}^{K} a_i w_i + b)$, where $\sigma(z) = \dfrac{1}{1 + \exp(-z)}$





sigmoid function

Training neural networks

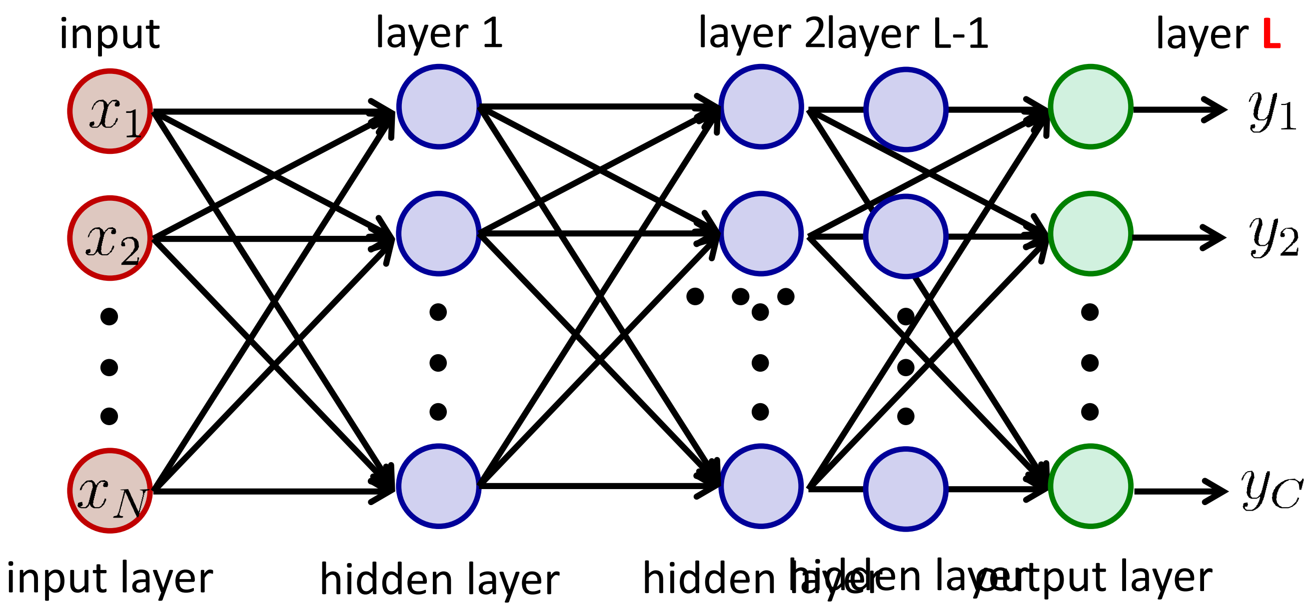
- Collect a set of labeled training data $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$
- Training neural networks: Finding network parameters $\theta = \{\mathbf{w}, \mathbf{b}\}$ to minimize the loss between true training label $\mathbf{y}_i$ and the estimated label, e.g.,

$$L(\theta) = \sum_{i=1}^{N} \|\mathbf{y}_i - g_{\mathbf{w}}(\mathbf{x}_i)\|^2$$

- Minimization can be done by gradient descent if $L(\cdot)$ is differentiable with respect to θ
- Back-propagation: a widely used method for optimizing multi-layer neural networks

What is deep neural networks (DNN)

- DNN is neural networks with many hidden layers



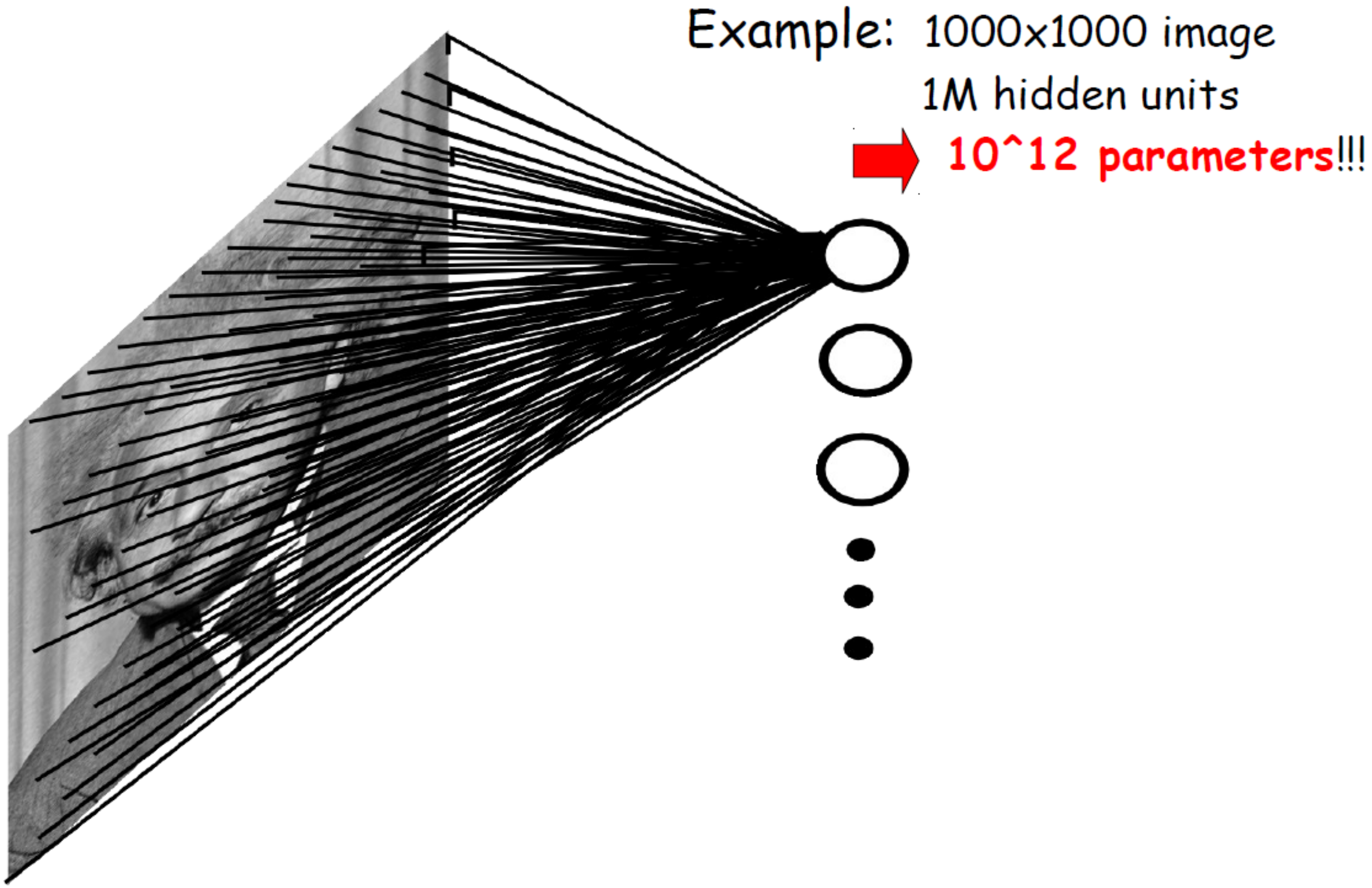
Outline

- Convolutional neural networks (CNN)
  - Conventional approaches vs. deep learning
  - Neural networks
  - Convolutional neural networks

- Representative CNN models

- CNN-based computer vision applications

of parameters in fully connected NN

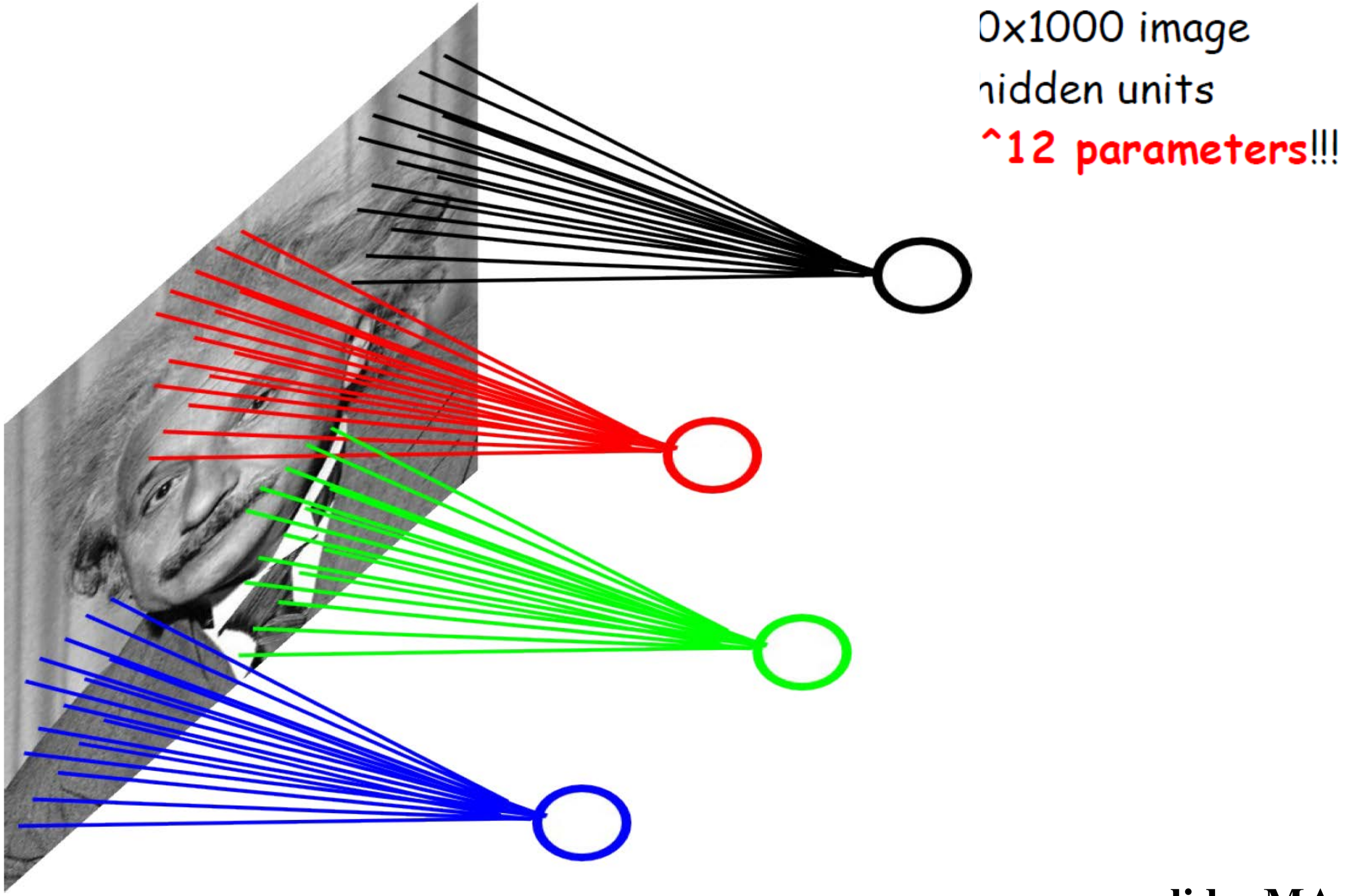


slide: MA Ranzato

Convolutional neural networks (CNN)

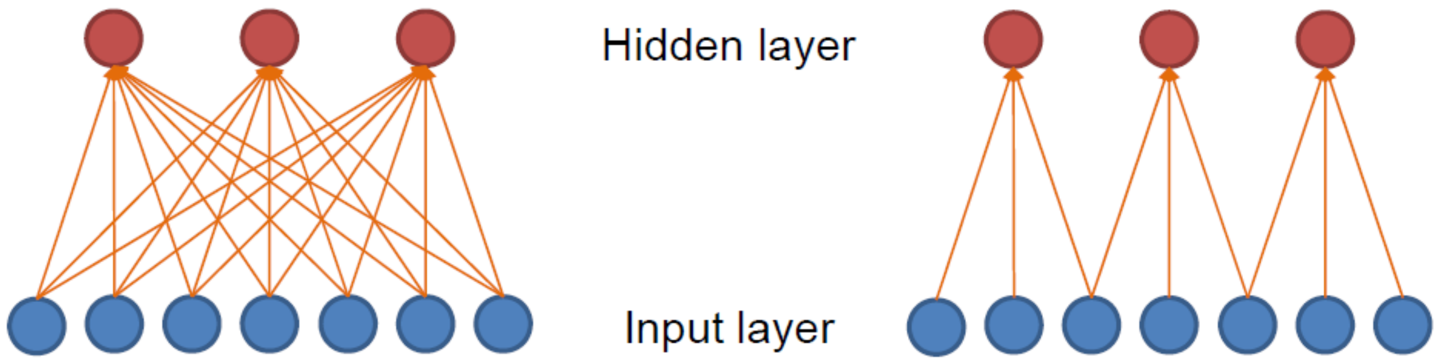
- CNN: a multi-layer neural network with
  1. Local connectivity
  2. Weight sharing
- Why local connectivity?
  - Spatial correlation is local (locality of spatial dependencies)
  - Reduce # of parameters
- Why weight sharing?
  - Statistics is at different locations (stationarity of statistics)
  - Reduce # of parameters

of parameters in fully connected NN



slide: MA Ranzato

CNN: Local connectivity

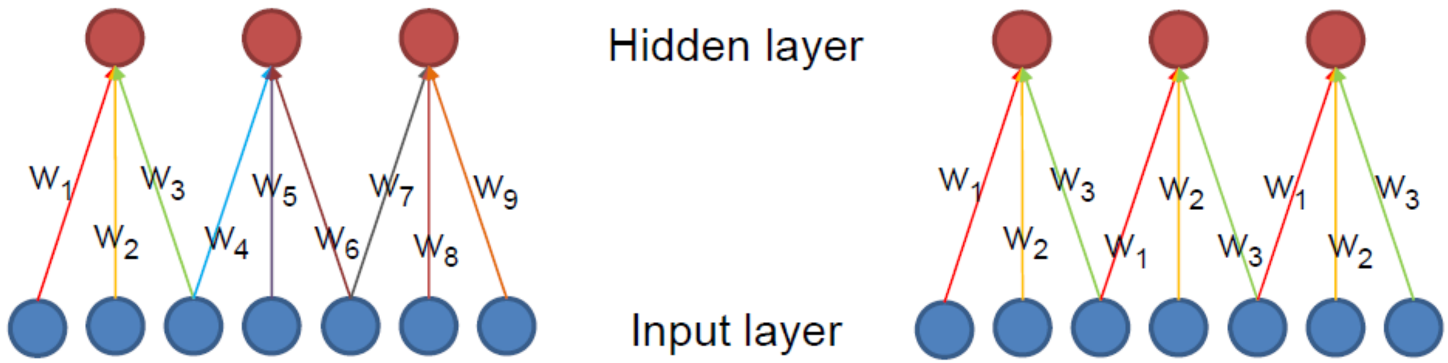


Global connectivity **Local** connectivity

- # input units (neurons): 7
- # hidden units: 3
- Number of parameters
  - Global connectivity: 3 x 7 = 21
  - Local connectivity: 3 x 3 = 9

slide: J.-B. Huang

CNN: Weight sharing



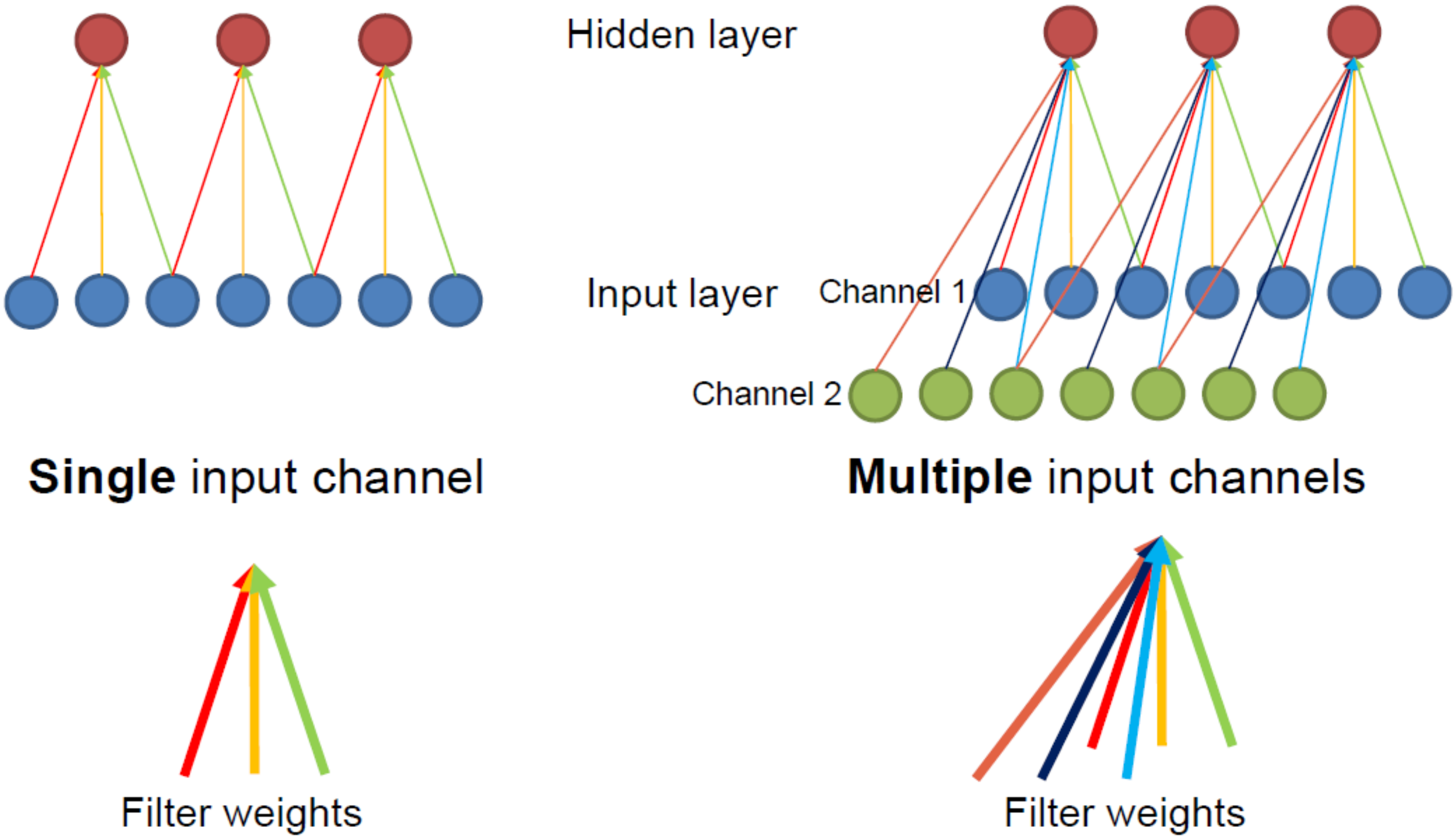
Without weight sharing

With weight sharing

- # input units (neurons): 7
- # hidden units: 3
- Number of parameters
  - Without weight sharing: 3 x 3 = 9
  - With weight sharing : 3 x 1 = 3

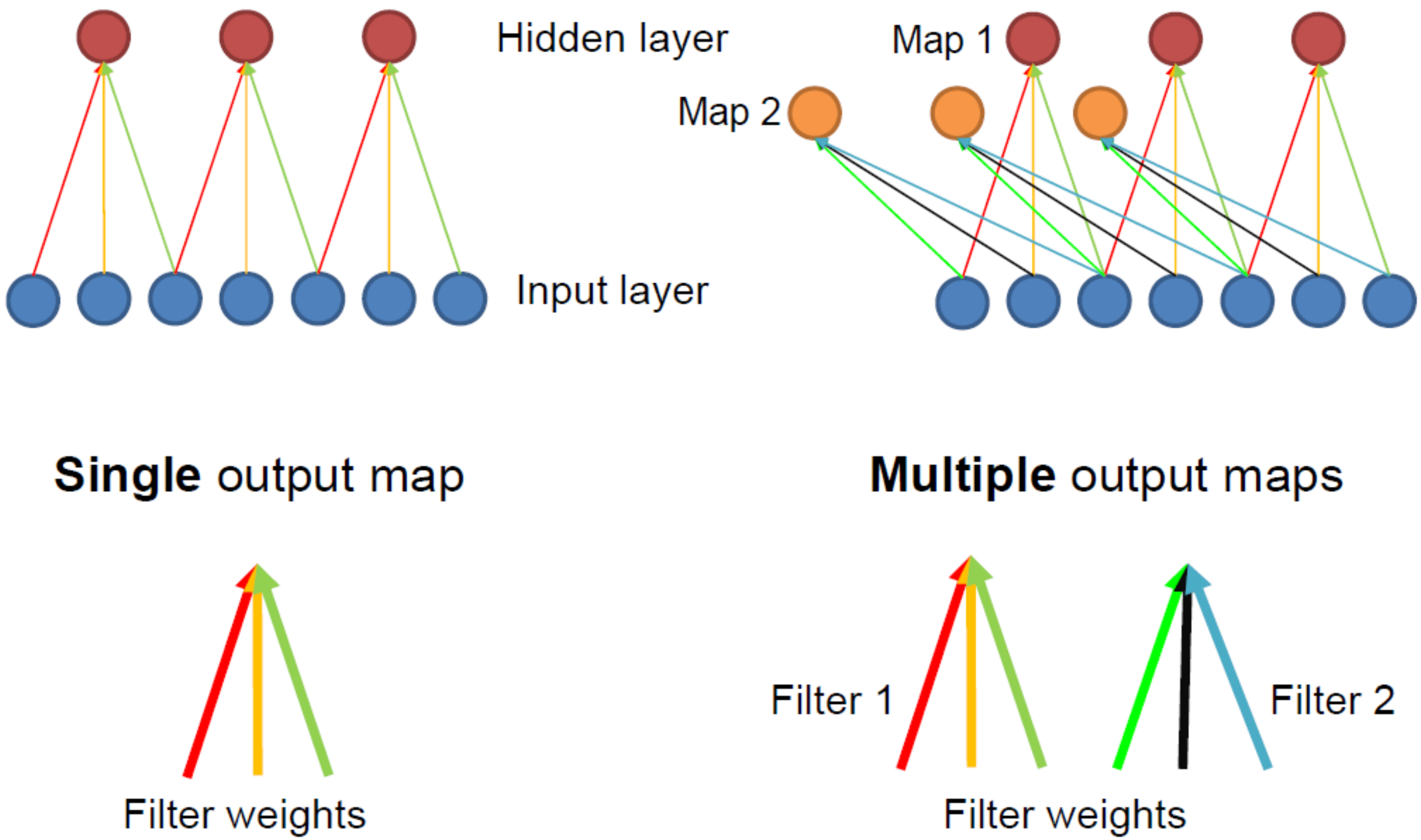
slide: J.-B. Huang

CNN with multiple input channels



slide: J.-B. Huang

CNN with multiple output channels



slide: J.-B. Huang

Putting them together

- Local connectivity
- Weight sharing
- Handling multiple input channels
- Handling multiple output maps

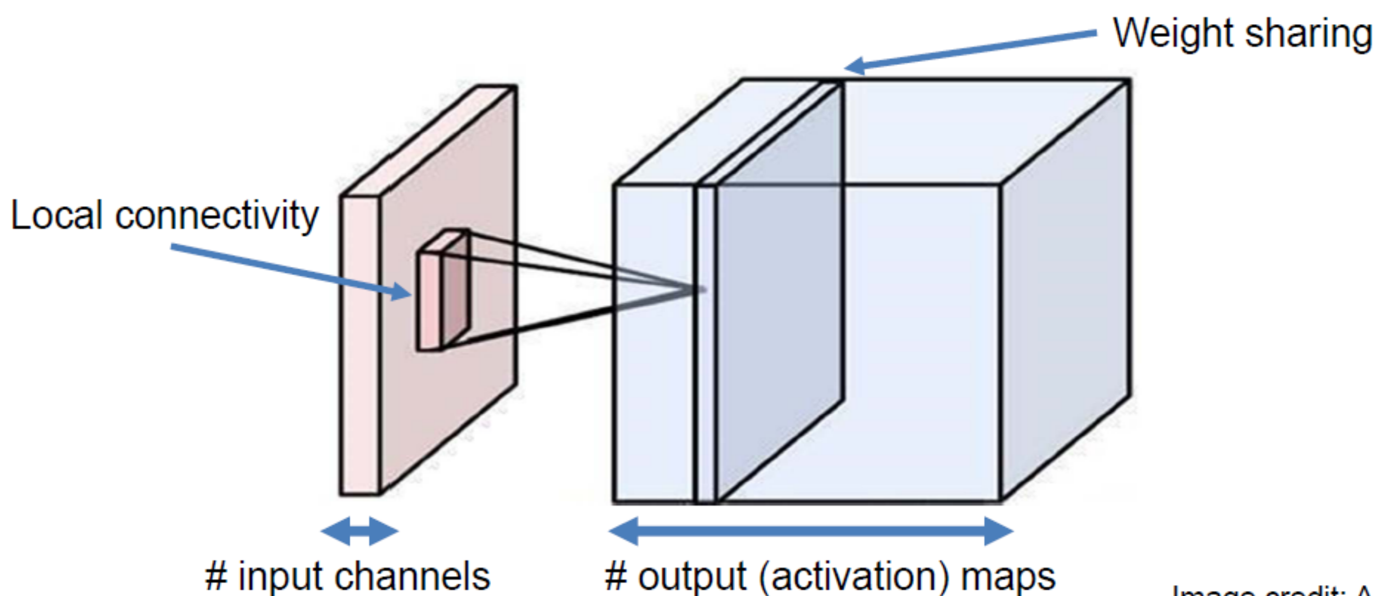
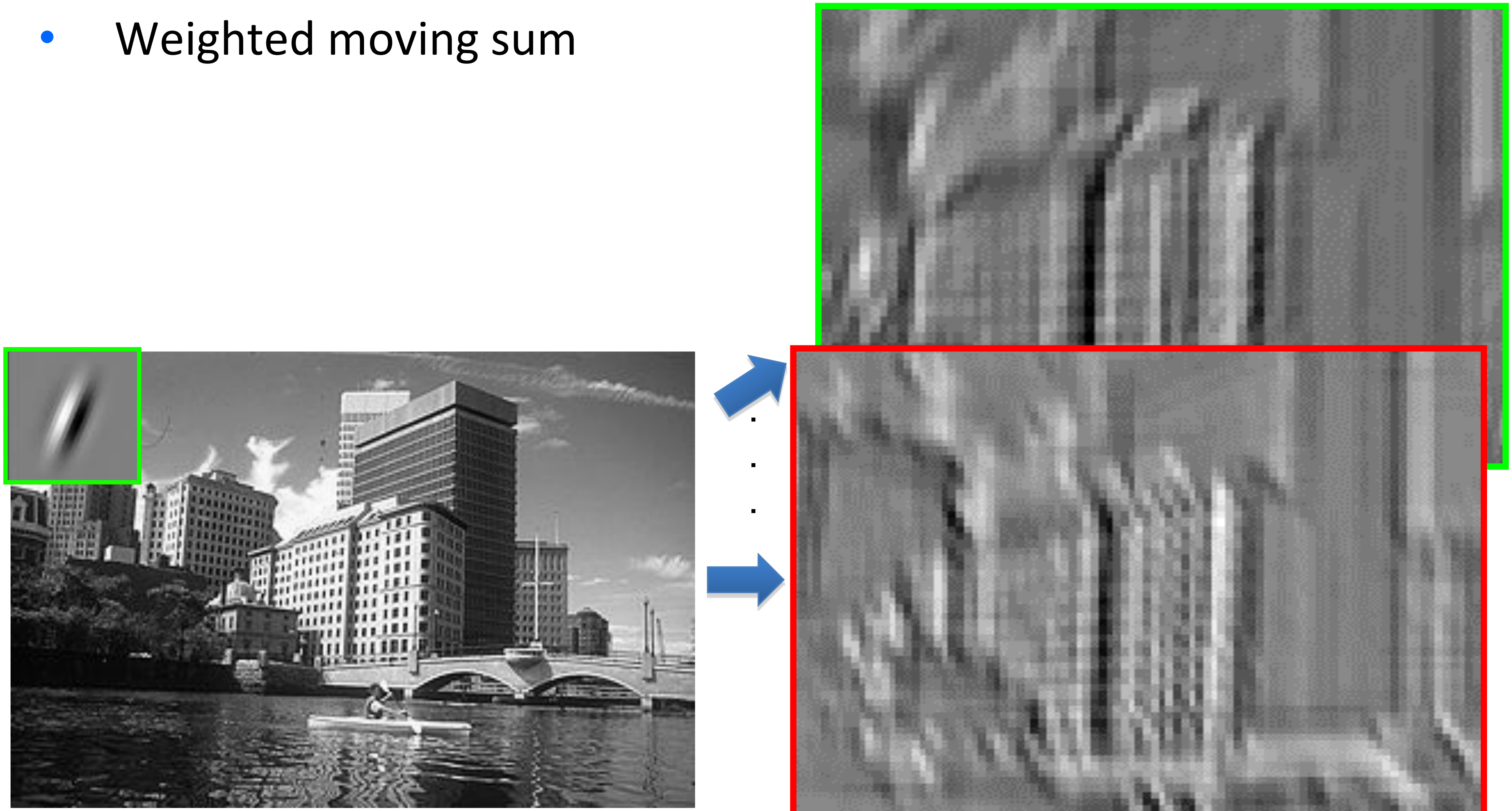


Image credit: A. Karpathy

slide: J.-B. Huang

What is a Convolution?

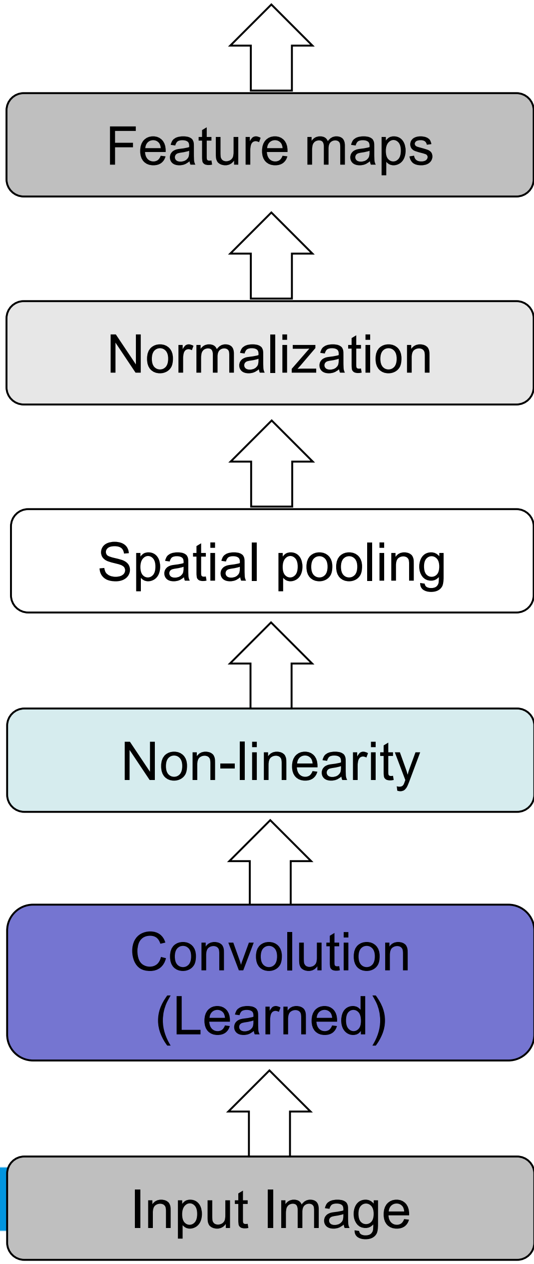
- Weighted moving sum



Input

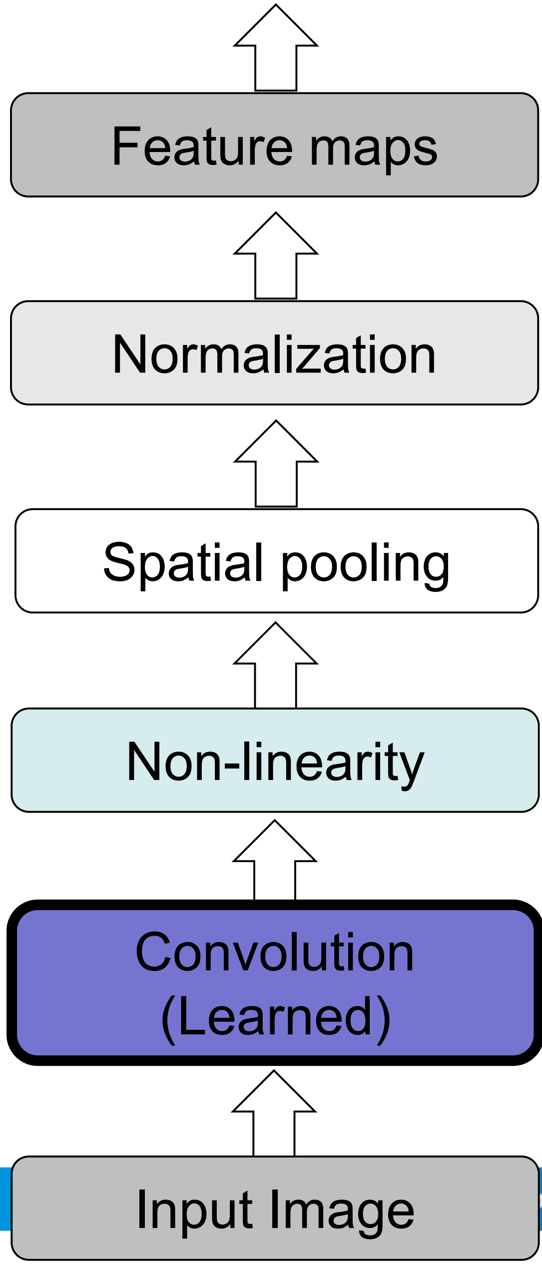
Feature Activation Map

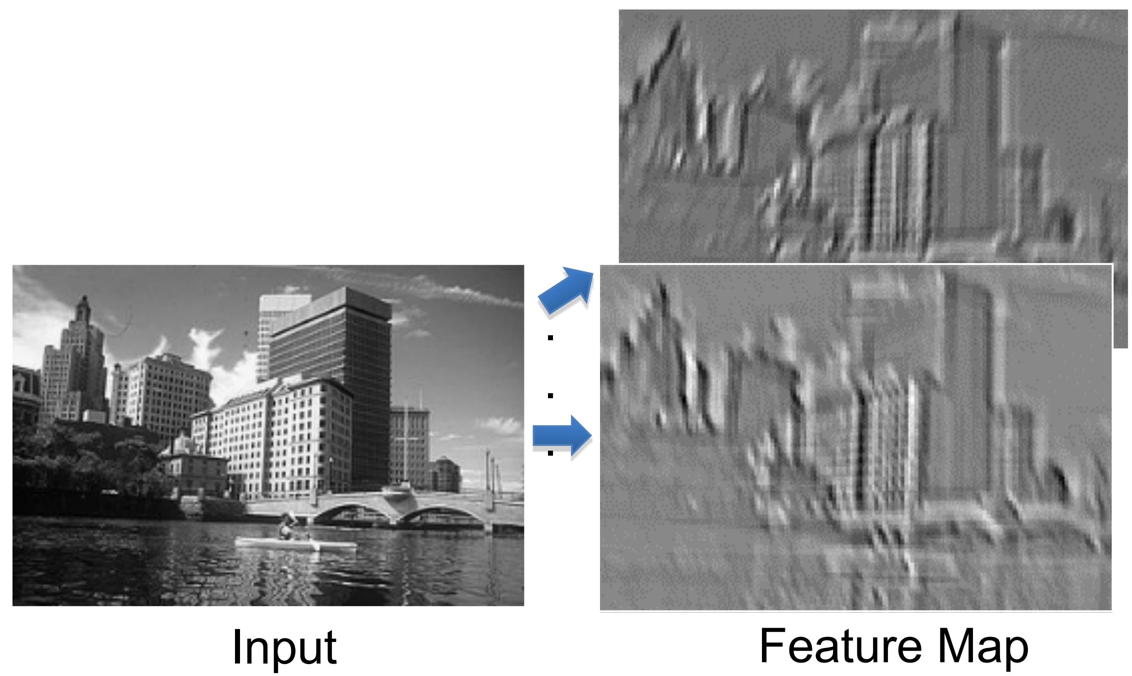
Convolutional Neural Networks



slide: S. Lazebnik

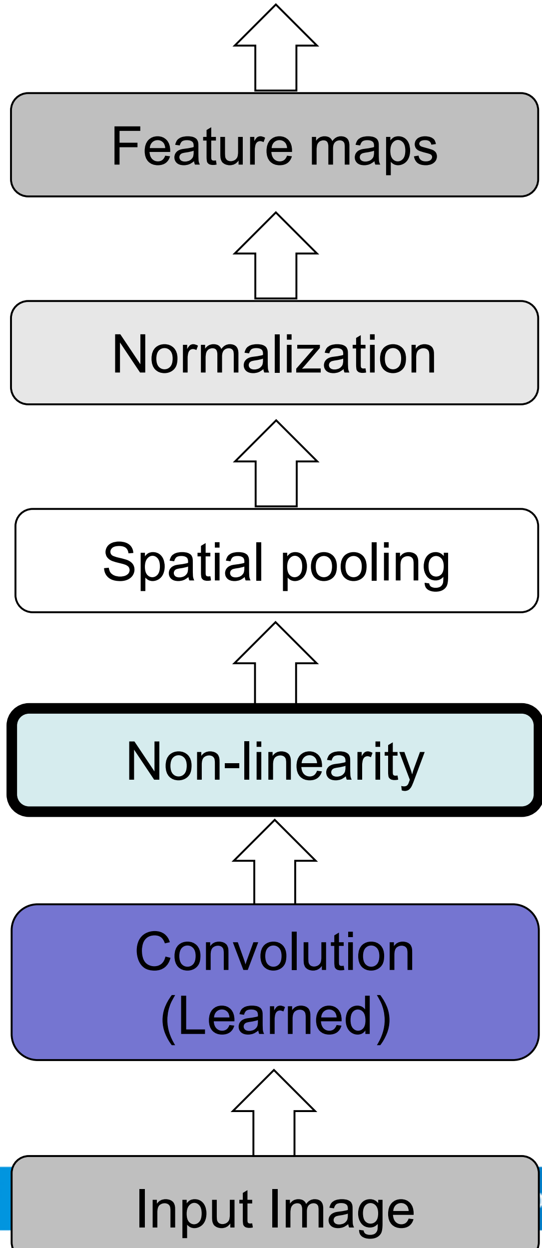
Convolutional Neural Networks



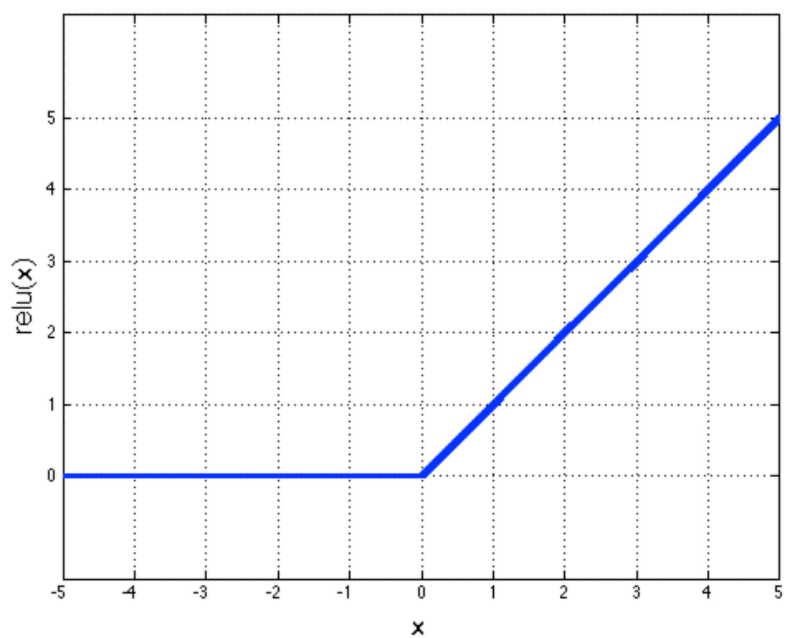


slide: S. Lazebnik

Convolutional Neural Networks

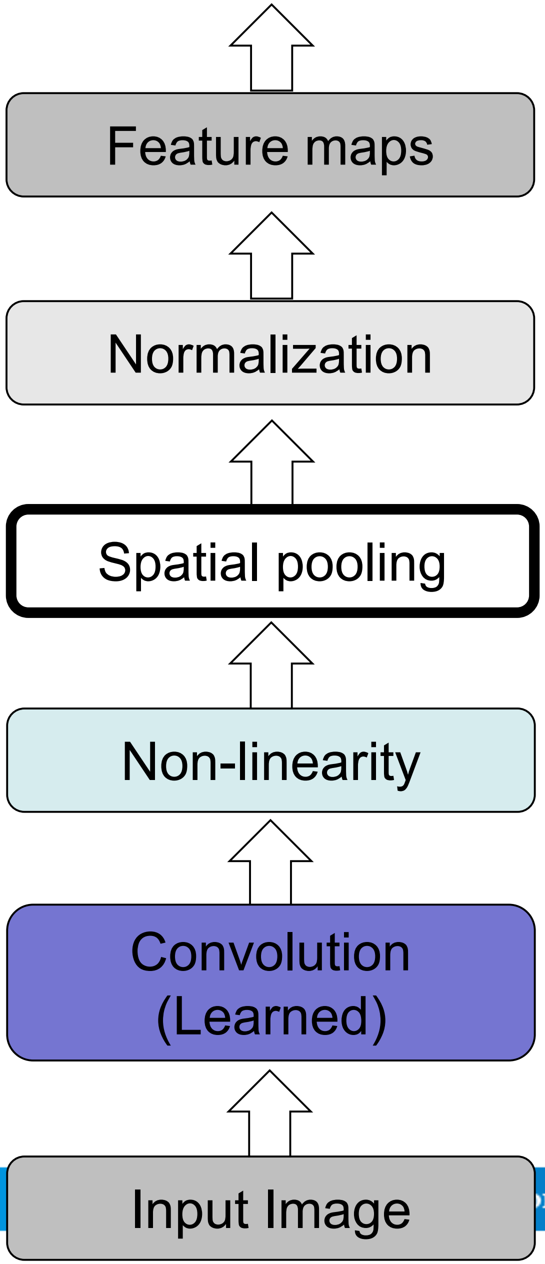


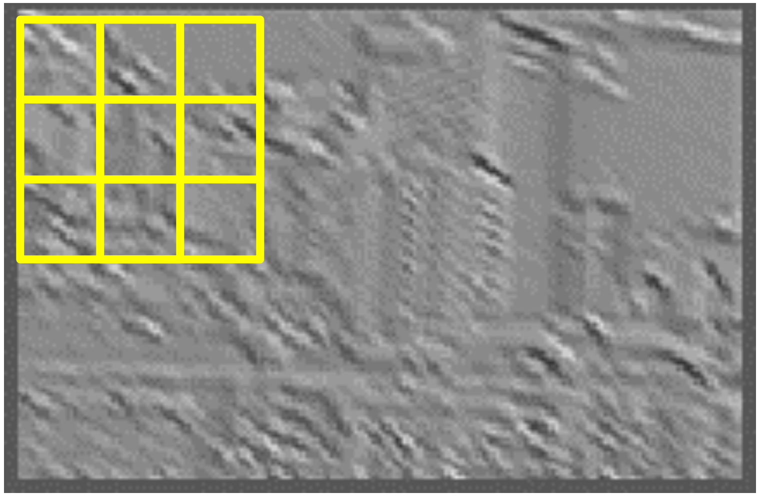
Rectified Linear Unit (ReLU)



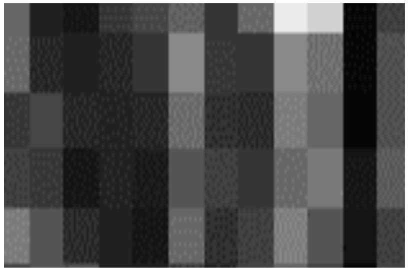
slide: S. Lazebnik

Convolutional Neural Networks





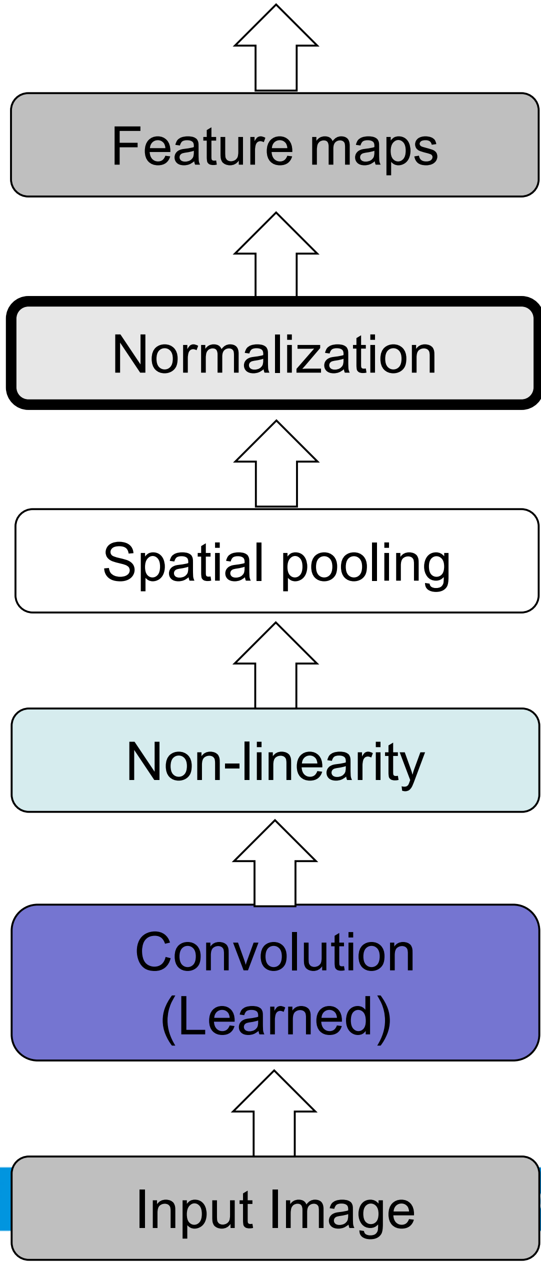
Max pooling

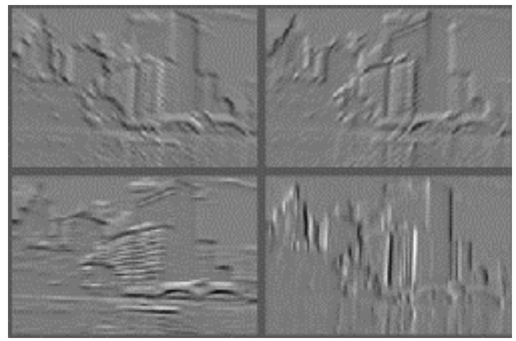


Max-pooling: a non-linear down-sampling

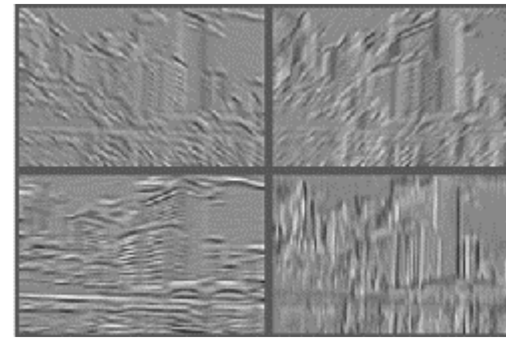
slide: S. Lazebnik

Convolutional Neural Networks





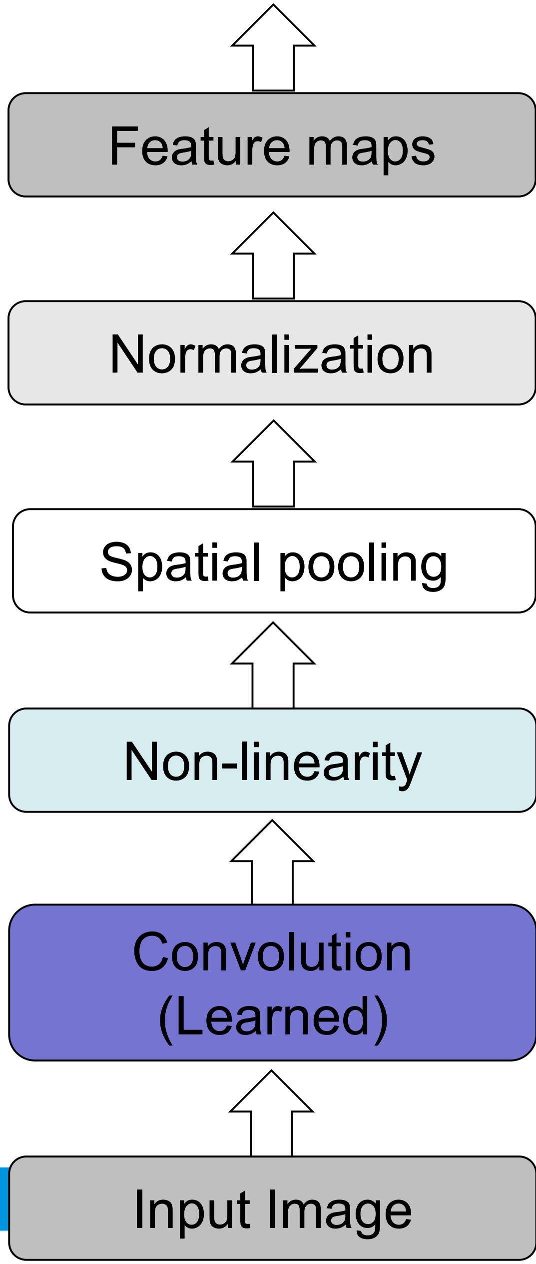
Feature Maps



Feature Maps After Contrast Normalization

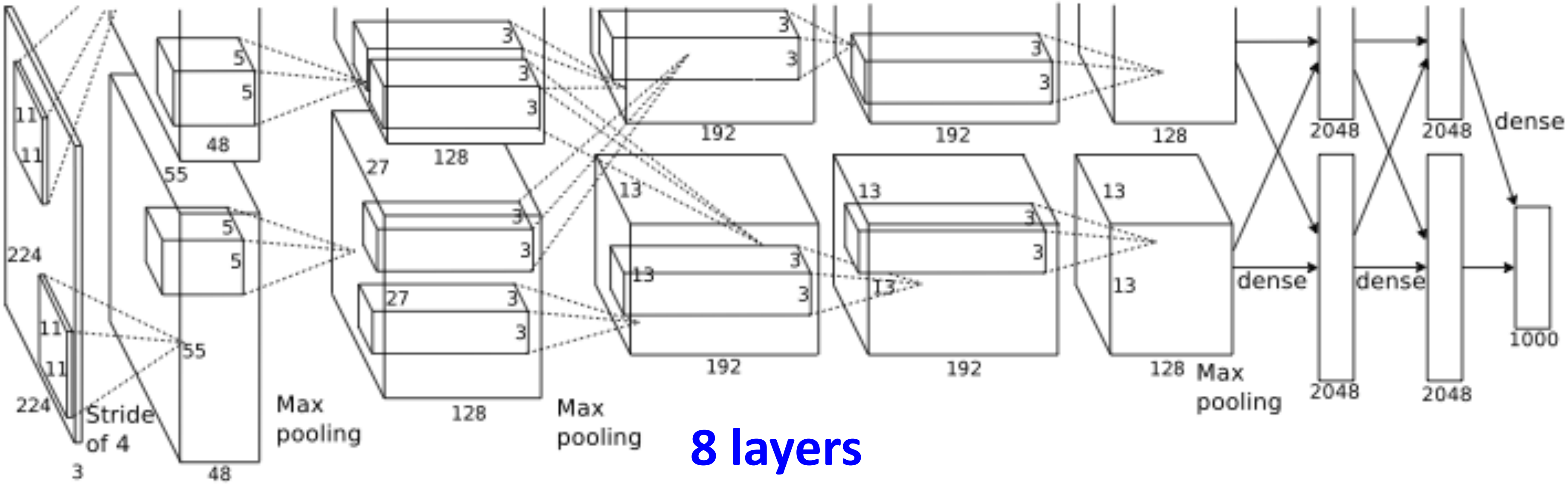
slide: S. Lazebnik

Convolutional Neural Networks



slide: S. Lazebnik

Modern CNN: AlexNet



8 layers

Input: 224*224*3=**150K**

Neurons: 290400+186624+64896+64896+43264+4096+4096+1000=**650K**

Weights: 11*11*3*48*2(35K)+5*5*48*128*2(307K)+128*3*3*192*4(884K)+ 192*3*3*192*2(663K)+192*3*3*128*2(442K)+6*6*128*2048*4(38M)+4096*4096(17M)+4096*1000(4M)=**60M**

- **More data (1.2M)**
- **Trained on two GPUs for a week**
- **Dropout**

slide: M. Sun

ImageNet ISLVRC 2012-2014: Object Recognition

Best non-convnet in 2012: 26.2%

Team	Year	Place	Error (top-5)	External data
SuperVision – Toronto (7 layers)	2012	-	16.4%	no
SuperVision	2012	1st	15.3%	ImageNet 22k
Clarifai – NYU (7 layers)	2013	-	11.7%	no
Clarifai	2013	1st	11.2%	ImageNet 22k
VGG – Oxford (16 layers)	2014	2nd	7.32%	no
GoogLeNet (19 layers)	2014	1st	6.67%	no
Human expert*			5.1%	

Team	Method	Error (top-5)
DeepImage - Baidu	Data augmentation + multi GPU	5.33%
PReLU-nets - MSRA	Parametric ReLU + smart initialization	4.94%
BN-Inception ensemble - Google	Reducing internal covariate shift	4.82%

slide: J.-B. Huang

Outline

- Convolutional neural networks (CNNs)
  - Conventional approaches vs. deep learning
  - Neural networks
  - Convolutional neural networks
- Representative CNN models
- CNN-based computer vision applications