

Category Recognition



Jia-Bin Huang

Virginia Tech

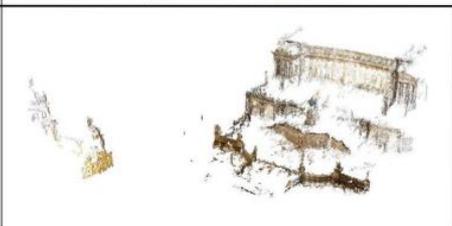
ECE 6554 Advanced Computer Vision

Administrative stuffs

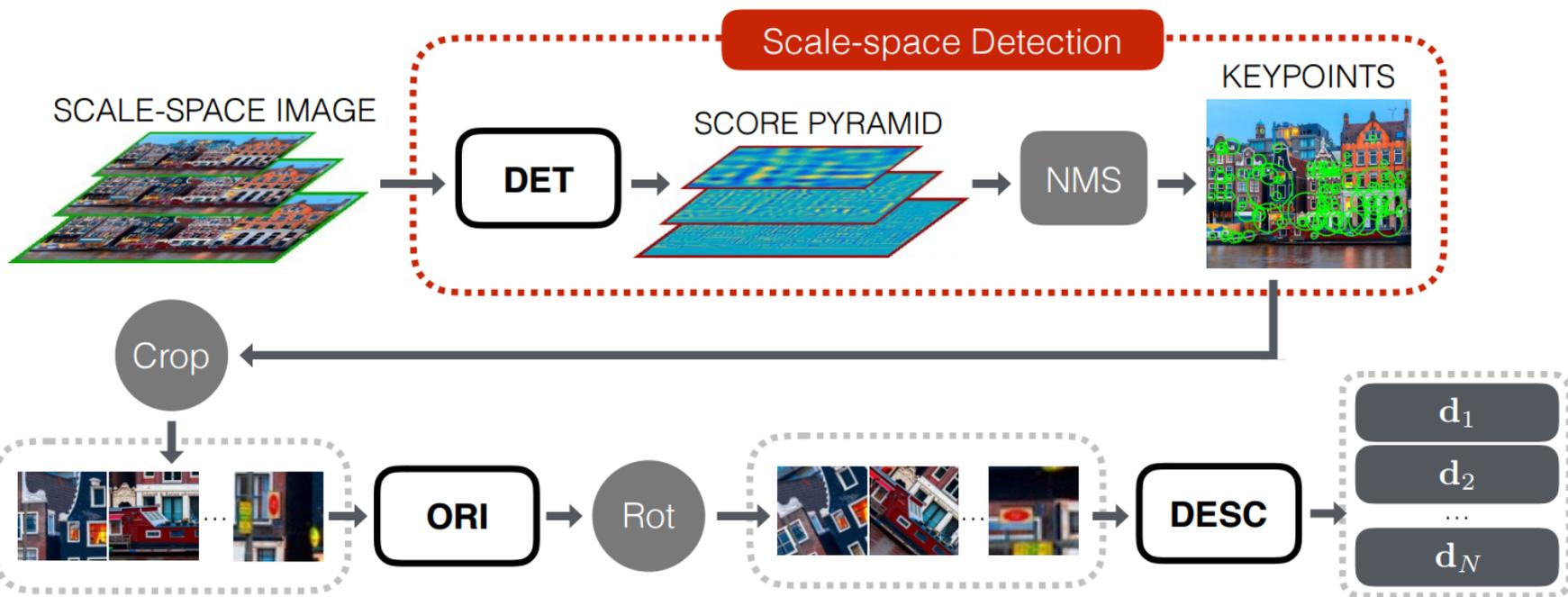
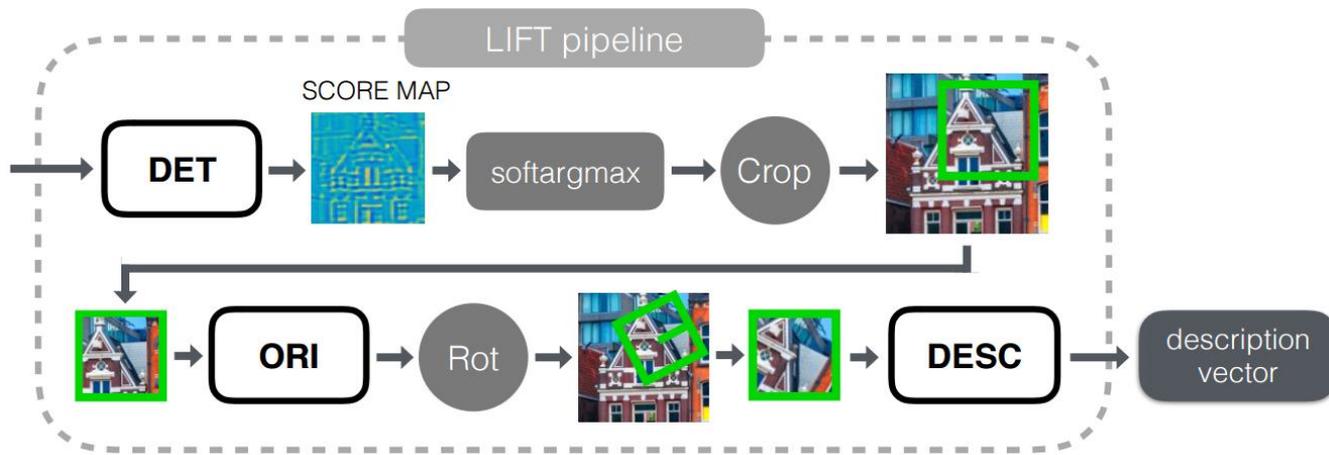
- Presentation and discussion leads assigned
 - <https://docs.google.com/spreadsheets/d/1P5pfyCio5flq3QCy4Mo1XS66I6d14jqDxE2Tny4efVs/edit#gid=0>
- Questions?

Today's class

- Finish instance recognition
- Category recognition
- Convolutional neural network

Day Image	Day Model	Night Model	Fused Night Model	Night Image
				
				
				
				

[From Dusk till Dawn: Modeling in the Dark, CVPR 2016](#)



Instance recognition

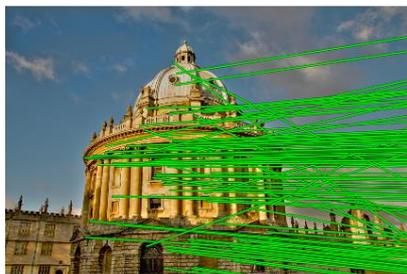
- Motivation – visual search
- Visual words
 - quantization, index, bags of words
- Spatial verification
 - affine; RANSAC, Hough
- Other text retrieval tools
 - tf-idf, query expansion
- Example applications

Instance recognition: remaining issues

- How to summarize the content of an entire image?
And gauge overall similarity?
- How large should the vocabulary be? How to perform quantization efficiently?
- Is having the same set of visual words enough to identify the object/scene? How to verify spatial agreement?
- How to score the retrieval results?

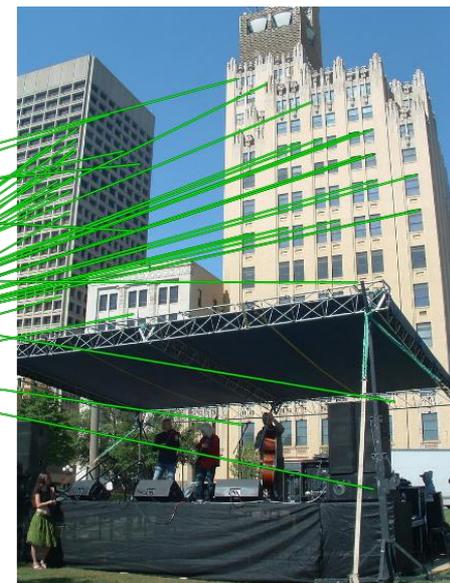
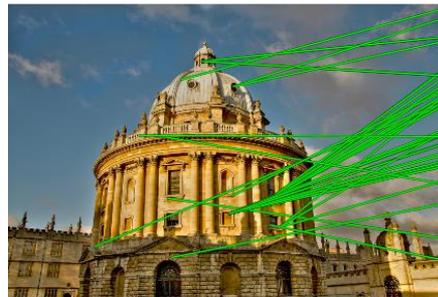
Spatial Verification

Query



DB image with high BoW similarity

Query



DB image with high BoW similarity

Both image pairs have many visual words in common.

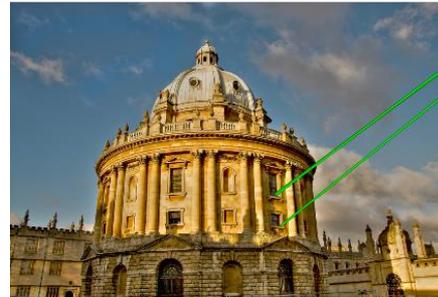
Spatial Verification

Query



DB image with high BoW similarity

Query



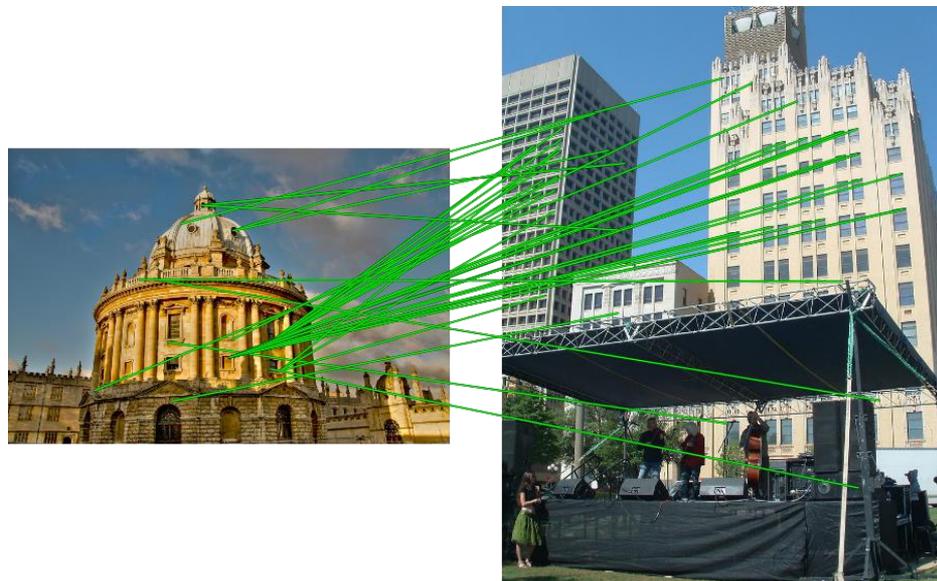
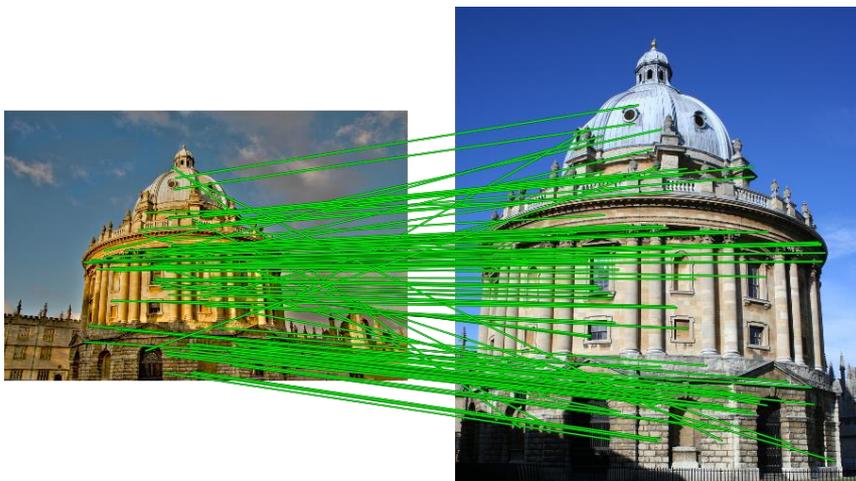
DB image with high BoW similarity

Only some of the matches are mutually consistent

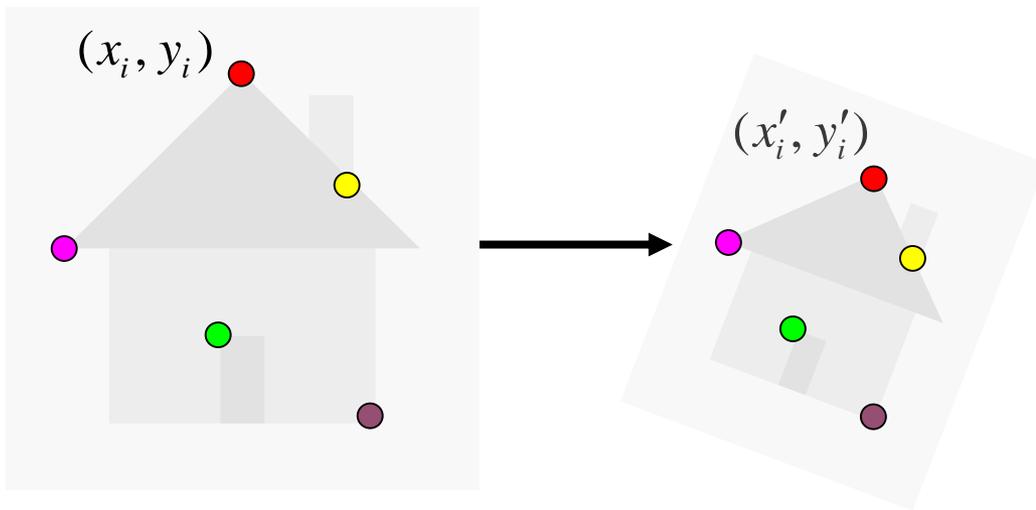
Spatial Verification: two basic strategies

- RANSAC
 - Typically sort by BoW similarity as initial filter
 - Verify by checking support (inliers) for possible transformations
 - e.g., “success” if find a transformation with $> N$ inlier correspondences
- Generalized Hough Transform
 - Let each matched feature cast a vote on location, scale, orientation of the model object
 - Verify parameters with enough votes

RANSAC verification



Recall: Fitting an affine transformation

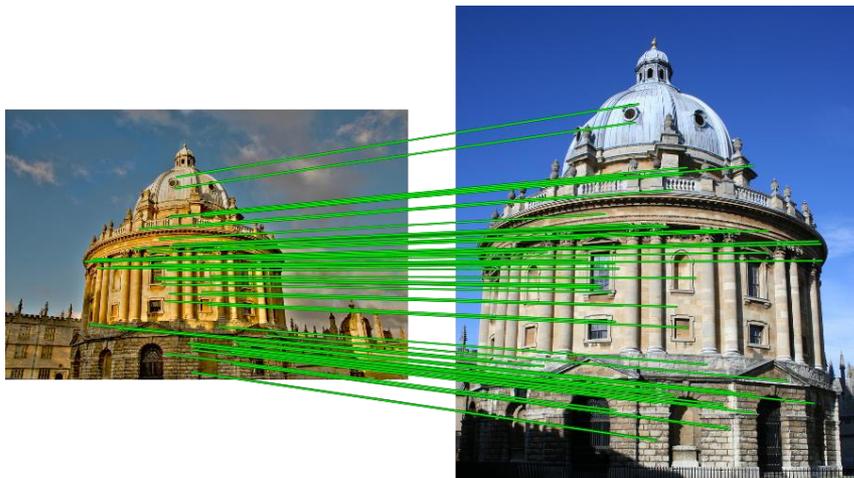
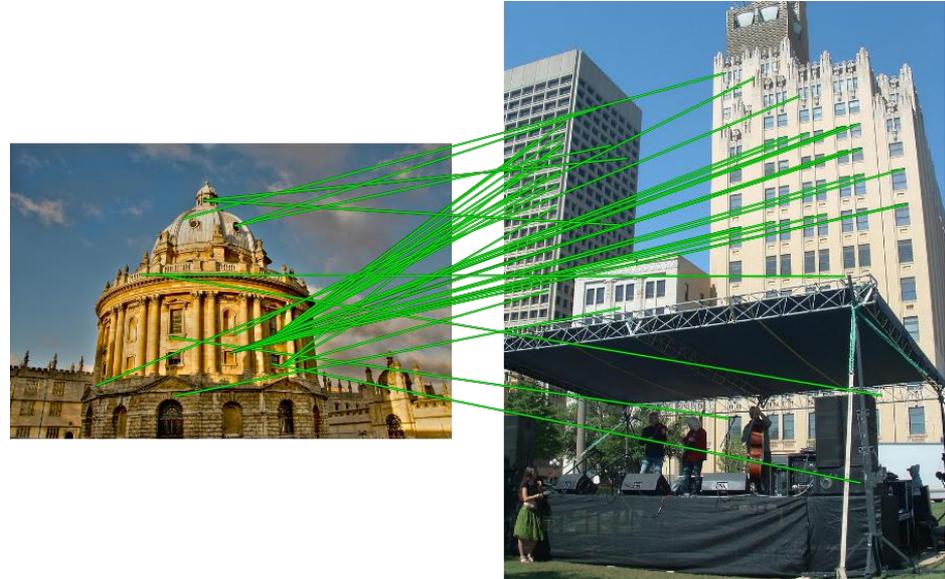
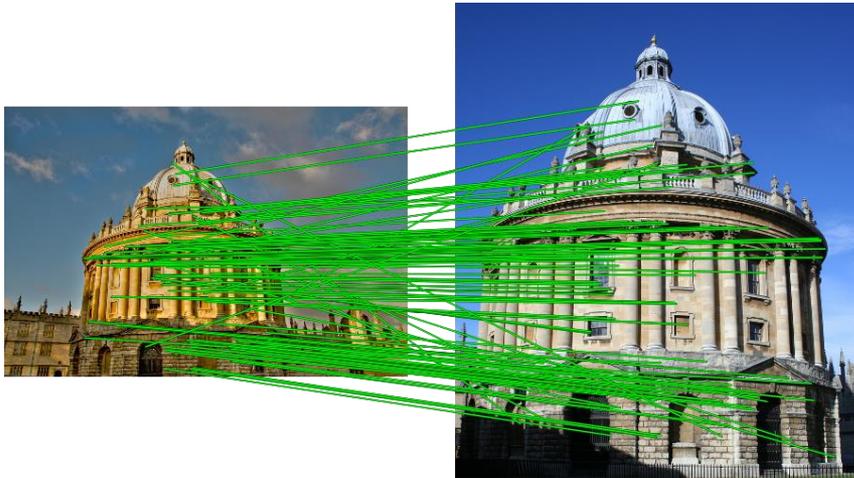


Approximates viewpoint changes for roughly planar objects and roughly orthographic cameras.

$$\begin{bmatrix} x'_i \\ y'_i \end{bmatrix} = \begin{bmatrix} m_1 & m_2 \\ m_3 & m_4 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} + \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

$$\begin{bmatrix} \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_i & y_i & 0 & 0 & 1 & 0 & m_1 \\ 0 & 0 & x_i & y_i & 0 & 1 & m_2 \\ \dots & \dots & \dots & \dots & \dots & \dots & m_3 \\ \dots & \dots & \dots & \dots & \dots & \dots & m_4 \\ \dots & \dots & \dots & \dots & \dots & \dots & t_1 \\ \dots & \dots & \dots & \dots & \dots & \dots & t_2 \end{bmatrix} = \begin{bmatrix} \dots \\ x'_i \\ y'_i \\ \dots \end{bmatrix}$$

RANSAC verification



Video Google System

1. Collect all words within query region
2. Inverted file index to find relevant frames
3. Compare word counts
4. Spatial verification

Sivic & Zisserman, ICCV 2003

- Demo online at :
<http://www.robots.ox.ac.uk/~vgg/research/vgoogle/index.html>

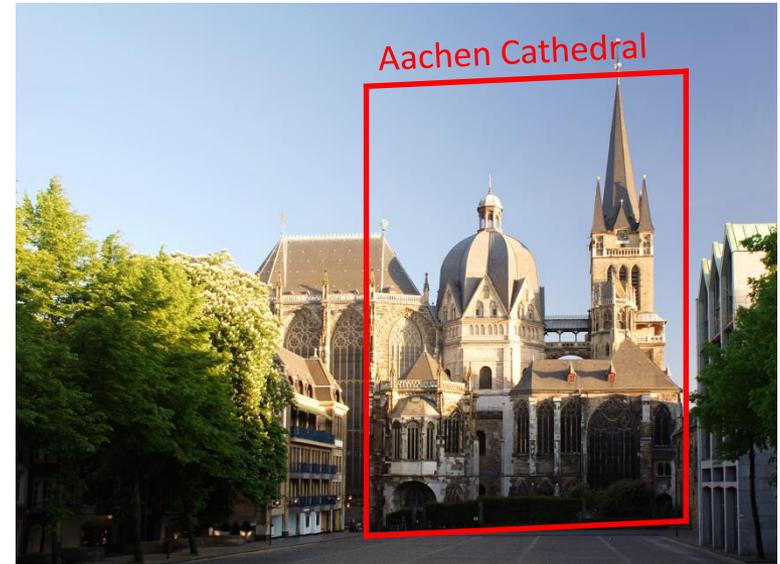


Query region



Retrieved frames

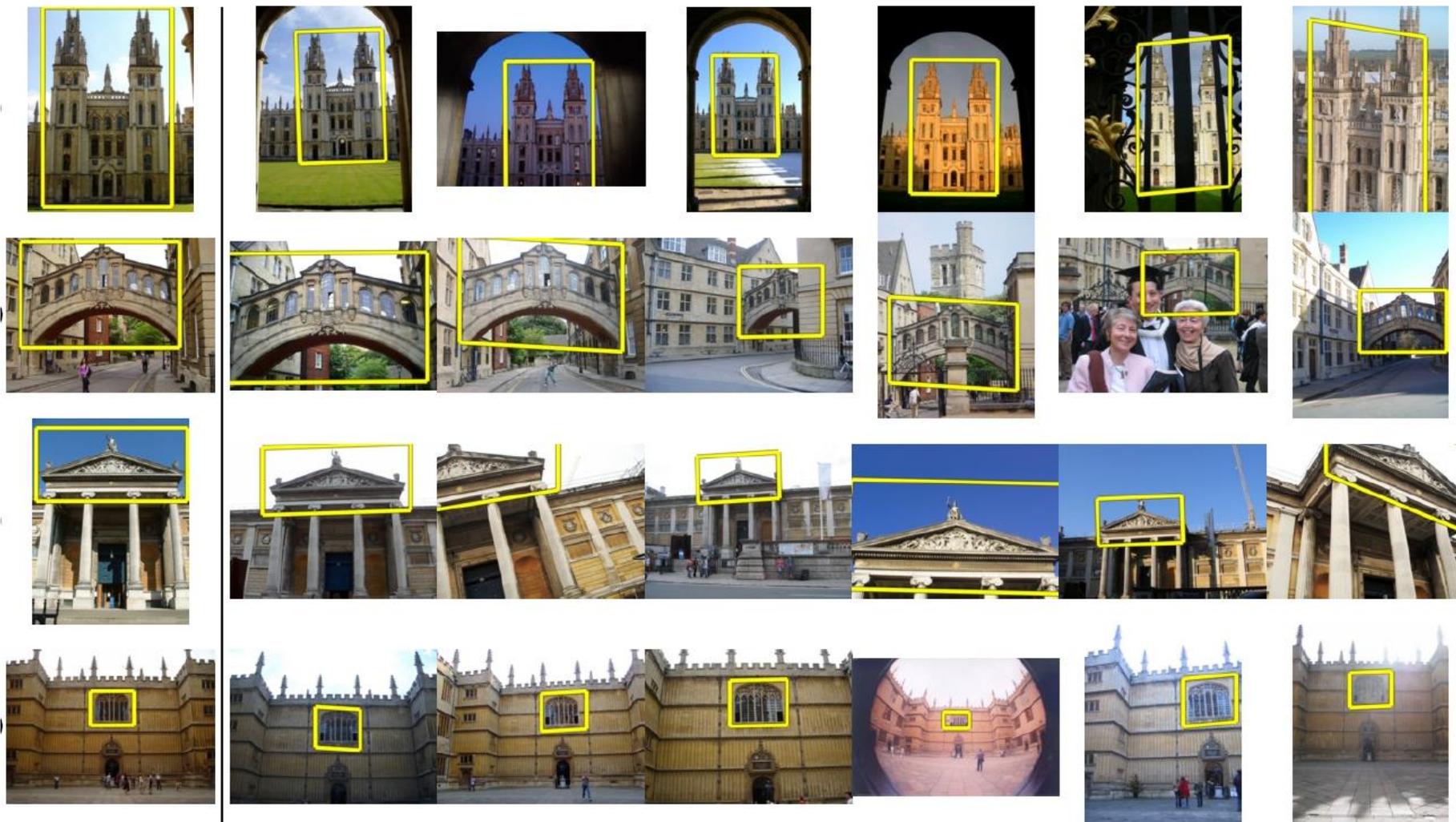
Example Applications



Mobile tourist guide

- Self-localization
- Object/building recognition
- Photo/video augmentation

Application: Large-Scale Retrieval



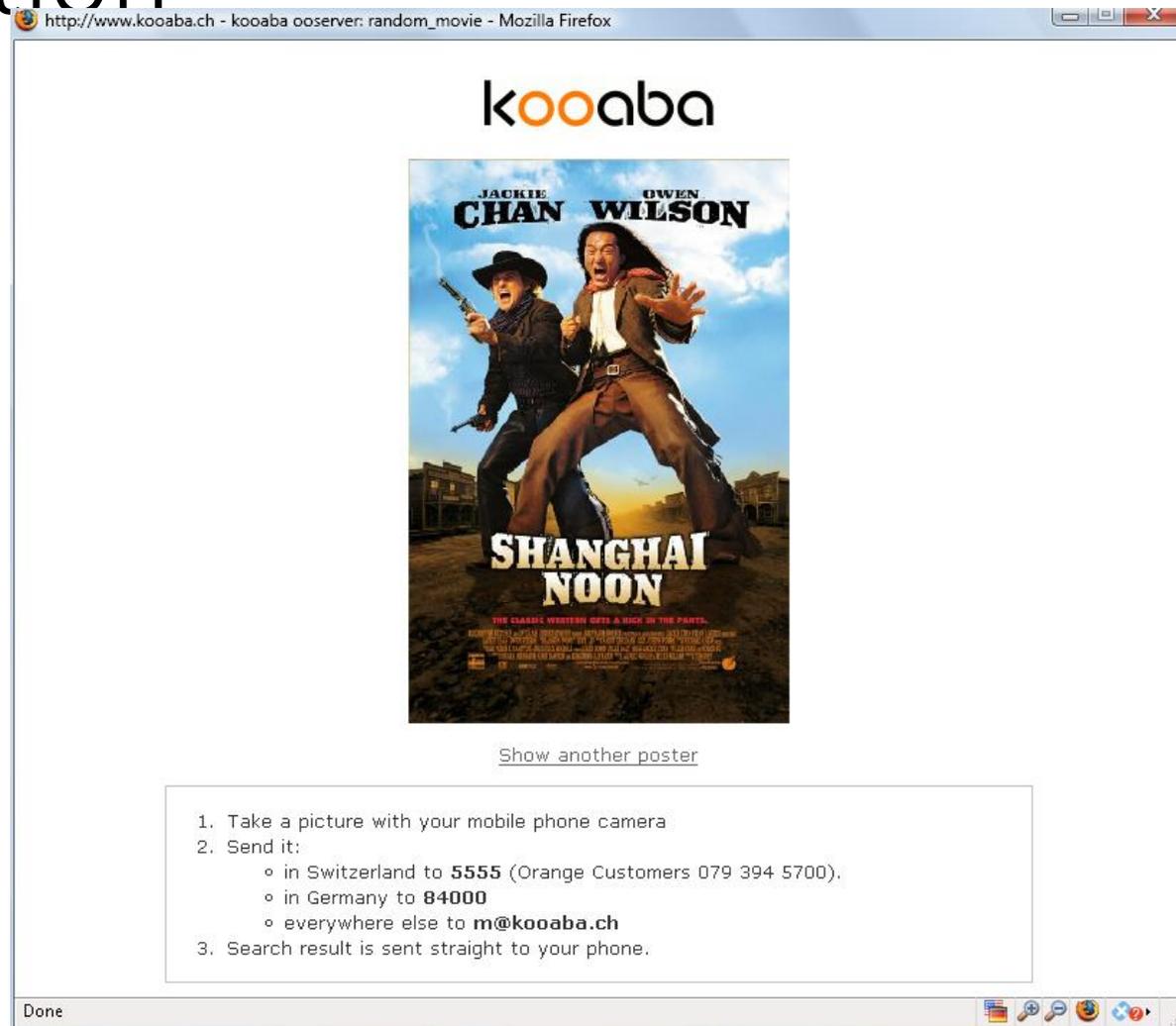
Query

Results from 5k Flickr images (demo available for 100k set)

Web Demo: Movie Poster Recognition

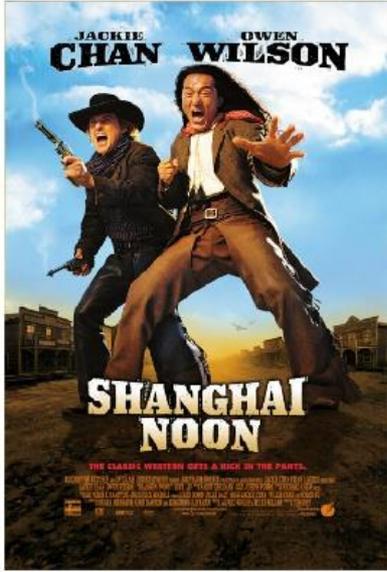
50'000 movie posters indexed

Query-by-image from mobile phone available in Switzerland



http://www.kooaba.ch - kooaba ooserver: random_movie - Mozilla Firefox

kooaba



SHANGHAI NOON

THE CLASSIC WESTERN GETS A HIGH IN THE PANTS.

Show another poster

1. Take a picture with your mobile phone camera
2. Send it:
 - in Switzerland to **5555** (Orange Customers 079 394 5700).
 - in Germany to **84000**
 - everywhere else to **m@kooaba.ch**
3. Search result is sent straight to your phone.

Done

http://www.kooaba.com/en/products_engine.html#



Google Goggles

Use pictures to search the web. [▶ Watch a video](#)



Get Google Goggles

Android (1.6+ required)

Download from [Android Market](#).

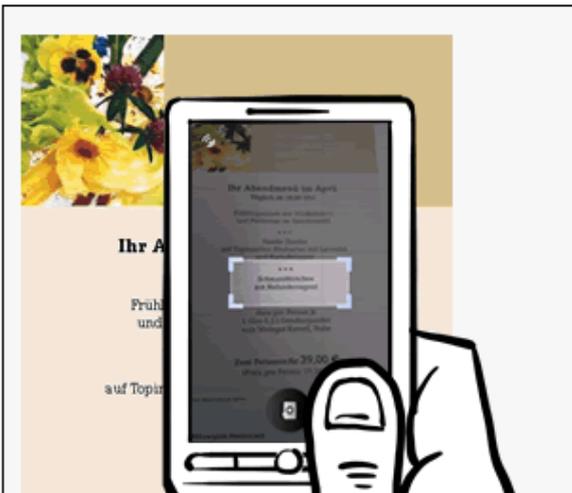
[Send Goggles to Android phone](#)

New! iPhone (iOS 4.0 required)

Download [from the App Store](#).

[Send Goggles to iPhone](#)

 New! Menu Crêpes-8 œufs-7						
Text	Landmarks	Books	Contact Info	Artwork	Wine	Logos

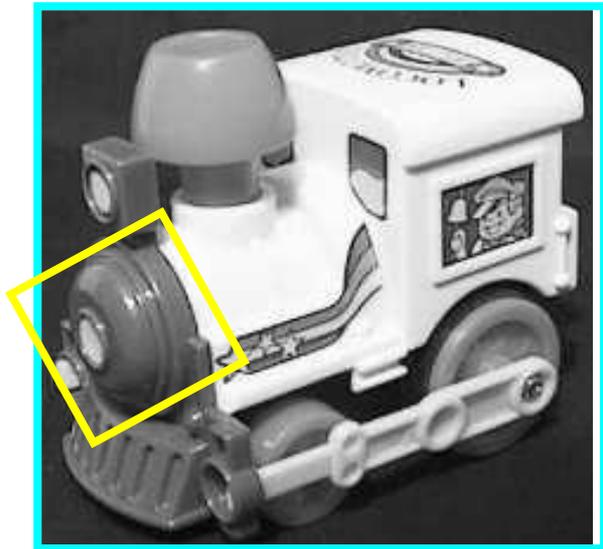


Spatial Verification: two basic strategies

- RANSAC
 - Typically sort by BoW similarity as initial filter
 - Verify by checking support (inliers) for possible transformations
 - e.g., “success” if find a transformation with $> N$ inlier correspondences
- Generalized Hough Transform
 - Let each matched feature cast a vote on location, scale, orientation of the model object
 - Verify parameters with enough votes

Voting: Generalized Hough Transform

- If we use scale, rotation, and translation invariant local features, then each feature match gives an alignment hypothesis (for scale, translation, and orientation of model in image).



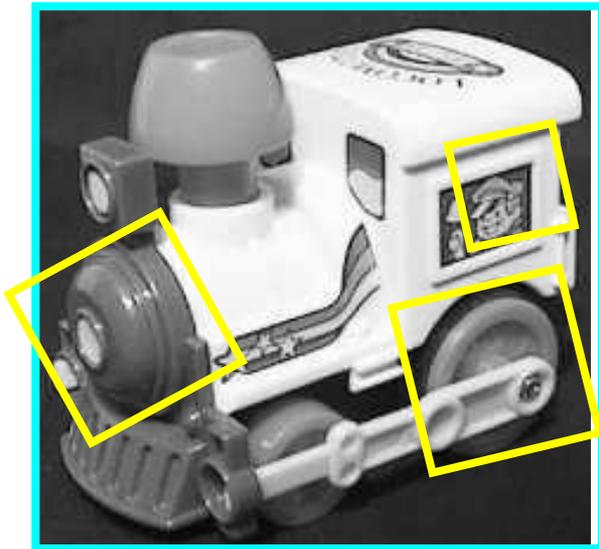
Model



Novel image

Voting: Generalized Hough Transform

- A hypothesis generated by a single match may be unreliable,
- So let each match **vote** for a hypothesis in Hough space



Model



Novel image

Gen Hough Transform details (Lowe's system)

- **Training phase:** For each model feature, record 2D location, scale, and orientation of model (relative to normalized feature frame)
- **Test phase:** Let each match btwn a test SIFT feature and a model feature vote in a 4D Hough space
 - Use broad bin sizes of 30 degrees for orientation, a factor of 2 for scale, and 0.25 times image size for location
 - Vote for two closest bins in each dimension
- Find all bins with at least three votes and perform geometric verification
 - Estimate least squares *affine* transformation
 - Search for additional features that agree with the alignment

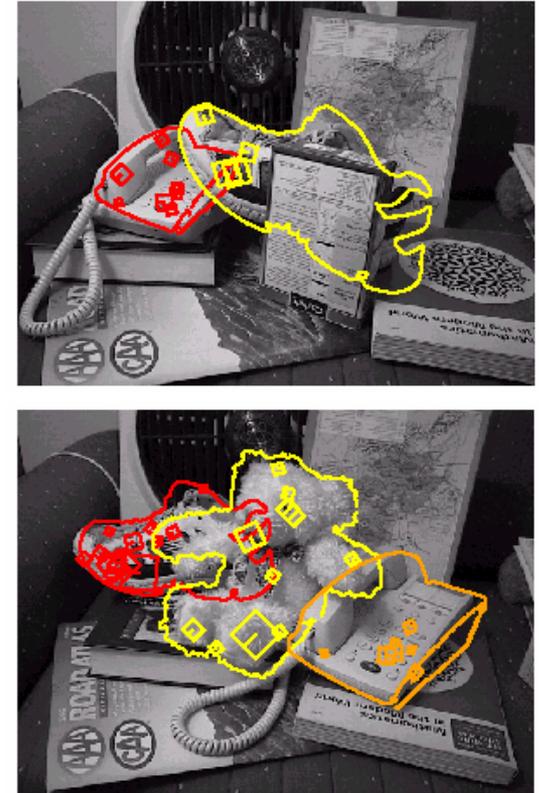
Example result



Background subtract for
model boundaries



Objects recognized,



Recognition in spite of
occlusion

Recall: difficulties of voting

- Noise/clutter can lead to as many votes as true target
- Bin size for the accumulator array must be chosen carefully
- In practice, good idea to make broad bins and spread votes to nearby bins, since verification stage can prune bad vote peaks.

Gen Hough vs RANSAC

GHT

- Single correspondence -> vote for all consistent parameters
- Represents uncertainty in the model parameter space
- Linear complexity in number of correspondences and number of voting cells; beyond 4D vote space impractical
- Can handle high outlier ratio

RANSAC

- Minimal subset of correspondences to estimate model -> count inliers
- Represents uncertainty in image space
- Must search all data points to check for inliers each iteration
- Scales better to high-d parameter spaces

What else can we borrow from text retrieval?

Index

"Along I-75," From Detroit to Florida; *inside back cover*
"Drive I-95," From Boston to Florida; *inside back cover*
1929 Spanish Trail Roadway; 101-102,104
511 Traffic Information; 83
A1A (Barrier Is) - I-95 Access; 86
AAA (and CAA); 83
AAA National Office; 88
Abbreviations,
 Colored 25 mile Maps; cover
 Exit Services; 196
 Travelogue; 85
Africa; 177
Agricultural Inspection Stns; 126
Ah-Tah-Thi-Ki Museum; 160
Air Conditioning, First; 112
Alabama; 124
Alachua; 132
 County; 131
Alafia River; 143
Alapaha, Name; 126
Alfred B MacLay Gardens; 106
Alligator Alley; 154-155
Alligator Farm, St Augustine; 169
Alligator Hole (definition); 157
Alligator, Buddy; 155
Alligators; 100,135,138,147,156
Anastasia Island; 170
Anhaica; 108-109,146
Apalachicola River; 112
Appleton Mus of Art; 136
Aquifer; 102
Arabian Nights; 94
Art Museum, Ringling; 147
Aruba Beach Cafe; 183
Aucilla River Project; 106
Babcock-Web WMA; 151
Bahia Mar Marina; 184
Baker County; 99
Barefoot Mailman; 182
Barge Canal; 137
Bee Line Expy; 80
Belz Outlet Mall; 89
Bernard Castro; 136
Big "I"; 165
Big Cypress; 155,158
Big Foot Monster; 105
Butterfly Center, McGuire; 134
CAA (see AAA)
CCC, The; 111,113,115,135,142
Ca d'Zan; 147
Caloosahatchee River; 152
 Name; 150
Canaveral Natnl Seashore; 173
Cannon Creek Airpark; 130
Canopy Road; 106,169
Cape Canaveral; 174
Castillo San Marcos; 169
Cave Diving; 131
Cayo Costa, Name; 150
Celebration; 93
Charlotte County; 149
Charlotte Harbor; 150
Chautauqua; 116
Chipley; 114
 Name; 115
Choctawatchee, Name; 115
Circus Museum, Ringling; 147
Citrus; 88,97,130,136,140,180
CityPlace, W Palm Beach; 180
City Maps,
 Fl Lauderdale Expwys; 194-195
 Jacksonville; 163
 Kissimmee Expwys; 192-193
 Miami Expressways; 194-195
 Orlando Expressways; 192-193
 Pensacola; 26
 Tallahassee; 191
 Tampa-St. Petersburg; 63
 St. Augustine; 191
Civil War; 100,108,127,138,141
Clearwater Marine Aquarium; 187
Collier County; 154
Collier, Barron; 152
Colonial Spanish Quarters; 168
Columbia County; 101,128
Corkina Building Material; 165
Corkscrew Swamp, Name; 154
Cowboys; 95
Crab Trap II; 144
Cracker, Florida; 88,95,132
Croastown Expy; 11,35,98,143
Cuban Bread; 184
Dade Battlefield; 140
Dade, Maj. Francis; 139-140,161
Dania Beach Hurricane; 184
Driving Lanes; 85
Duval County; 163
Eau Gallie; 175
Edison, Thomas; 152
Eglin AFB; 116-118
Eight Reale; 176
Ellenton; 144-145
Emanuel Point Wreck; 120
Emergency Callboxes; 83
Epiphytes; 142,148,157,159
Escambia Bay; 119
 Bridge (I-10); 119
 County; 120
Esterro; 153
Everglade; 90,95,139-140,154-160
 Draining of; 156,181
 Wildlife MA; 160
 Wonder Gardens; 154
Falling Waters SP; 115
Fantasy of Flight; 95
Fayer Dykes SP; 171
Fires, Forest; 166
Fires, Prescribed ; 148
Fisherman's Village; 151
Flagler County; 171
Flagler, Henry; 97,165,167,171
Florida Aquarium; 186
Florida,
 12,000 years ago; 187
 Cavern SP; 114
 Map of all Expressways; 2-3
 Mus of Natural History; 134
 National Cemetery ; 141
 Part of Africa; 177
 Platform; 187
 Sheriff's Boys Camp; 126
 Sports Hall of Fame; 130
 Sun 'n Fun Museum; 97
 Supreme Court; 107
Florida's Turnpike (FTP); 178,189
25 mile Strip Maps; 66
 Administration; 189
 Coin System; 190
 Exit Services; 189
 HEFT; 76,161,190
 History; 189
 Names; 189
 Service Plazas; 190
 Spur SR91; 76

China is forecasting a trade surplus of \$90bn (£51bn) to \$100bn this year, a threefold increase on 2004's \$32bn. The Commerce Ministry said the surplus would be predicted 30% jump in exports with a 18% rise in imports. Further a rise in imports. China's deliberate the surplus one factor Xiaochua more to be stayed within value of the yuan July and permitted it to trade freely. However, Beijing has made that it will take its time and tread carefully allowing the yuan to rise further in value.

China, trade, surplus, commerce, exports, imports, US, yuan, bank, domestic, foreign, increase, trade, value

tf-idf weighting

- **T**erm frequency – **i**nverse **d**ocument frequency
- Describe frame by frequency of each word within it, downweight words that appear often in the database
- (Standard weighting for text retrieval)

Number of occurrences of word i in document d

Number of words in document d

$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i}$$

Total number of documents in database

Number of documents word i occurs in, in whole database

Query Expansion

Results

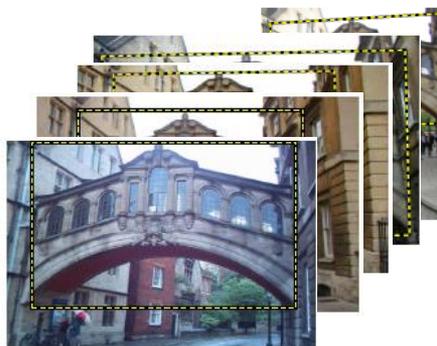


Query image

Spatial verification



New results



New query



Chum, Philbin, Sivic, Isard, Zisserman: Total Recall..., ICCV 2007

Slide credit: Ondrej Chum

Recognition via alignment

Pros:

- Effective when we are able to find reliable features within clutter
- Great results for matching specific instances

Cons:

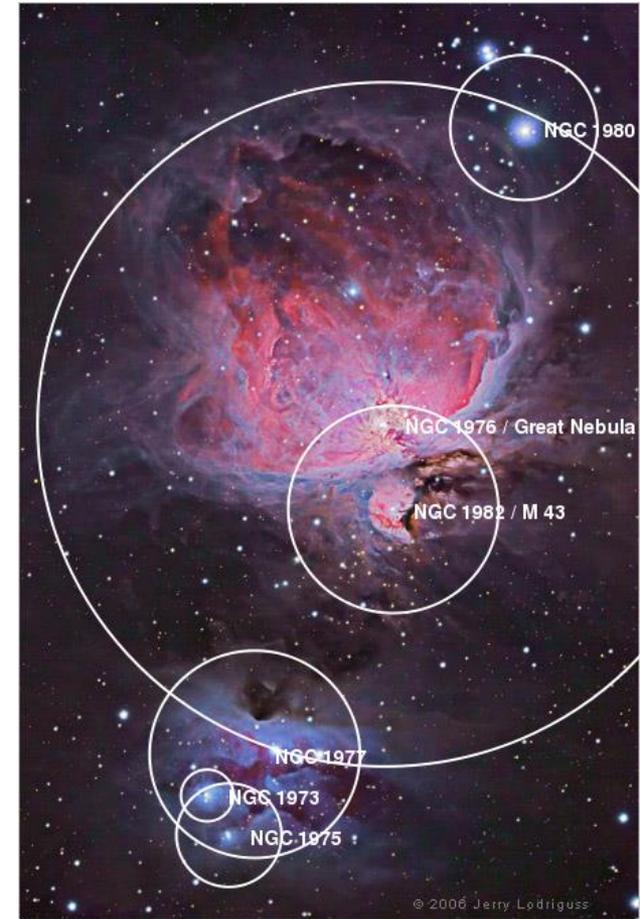
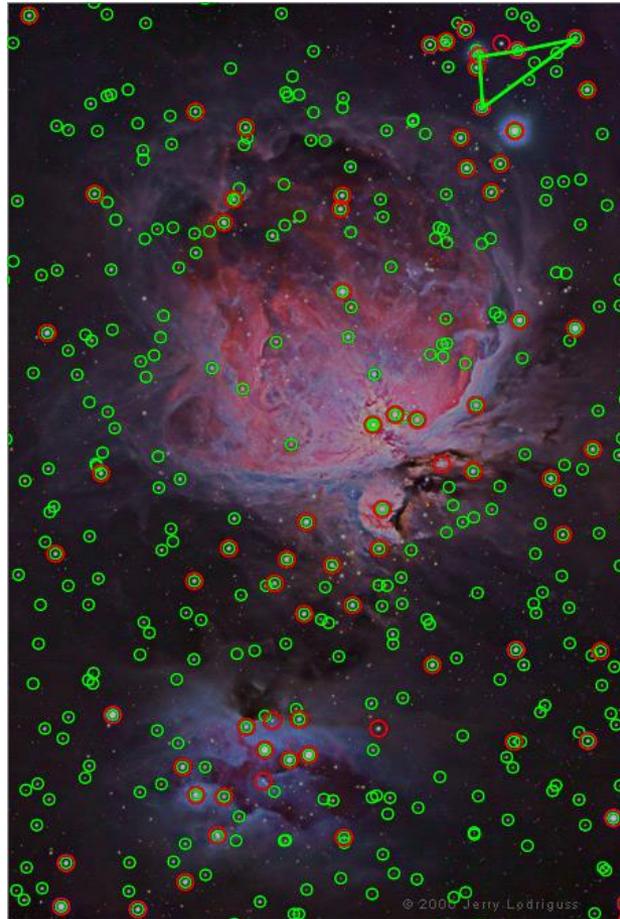
- Scaling with number of models
- Spatial verification as post-processing – not seamless, expensive for large-scale problems
- Not suited for category recognition.

Making the Sky Searchable: Fast Geometric Hashing for Automated Astrometry

Sam Roweis, Dustin Lang & Keir Mierle
University of Toronto

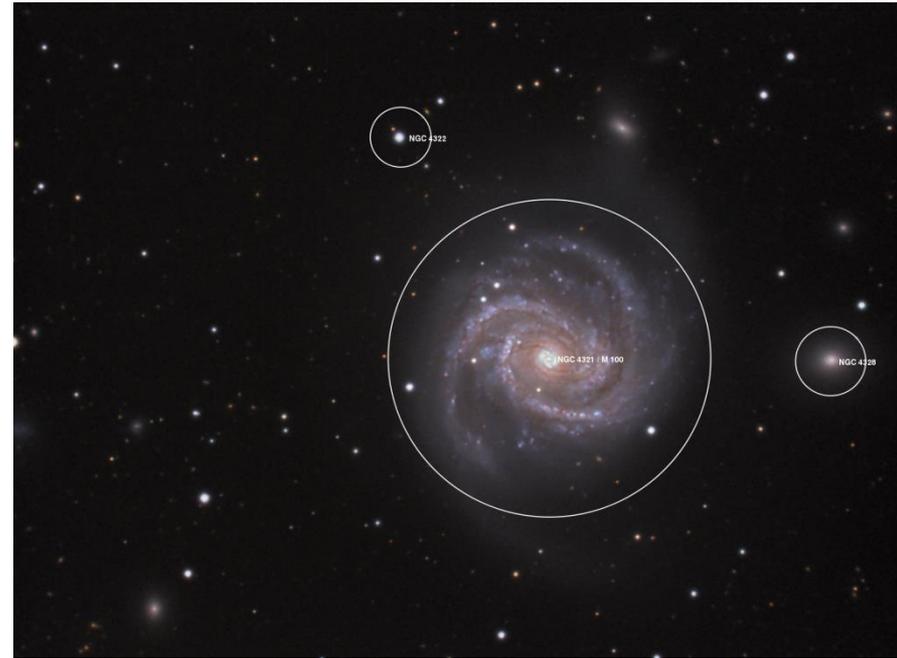
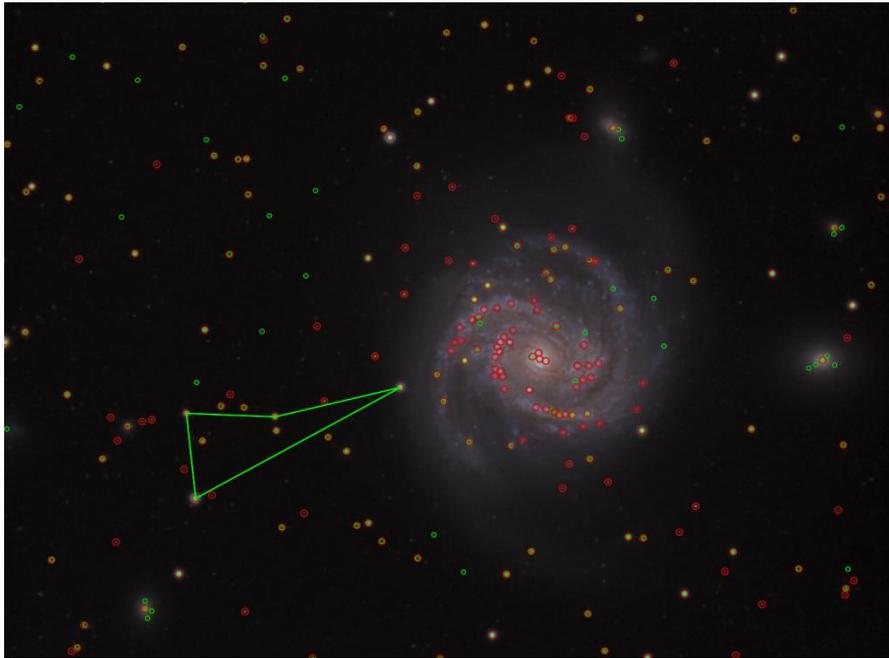
David Hogg & Michael Blanton
New York University

Example

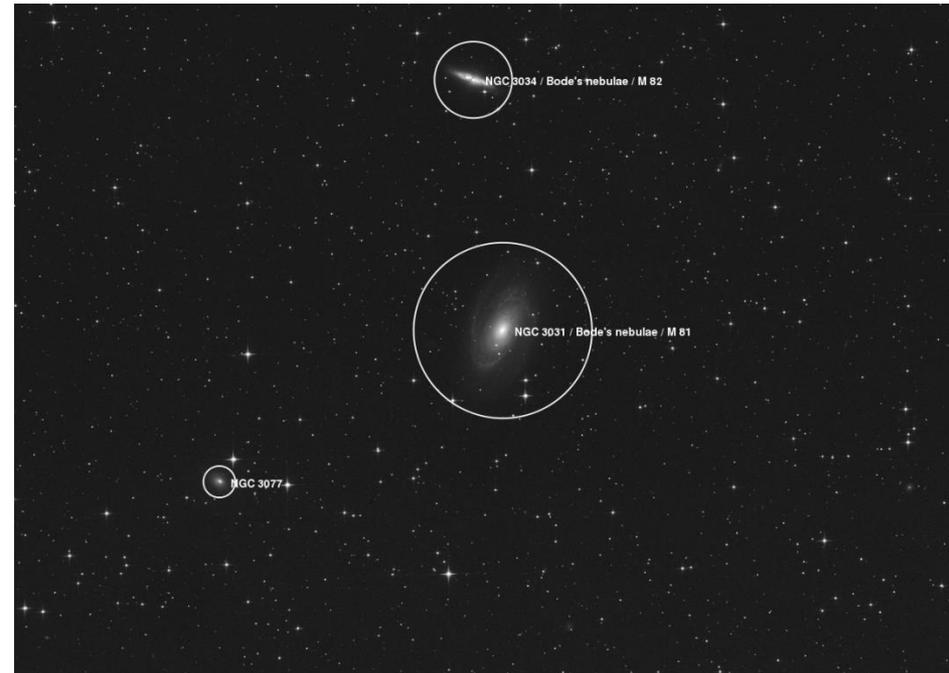
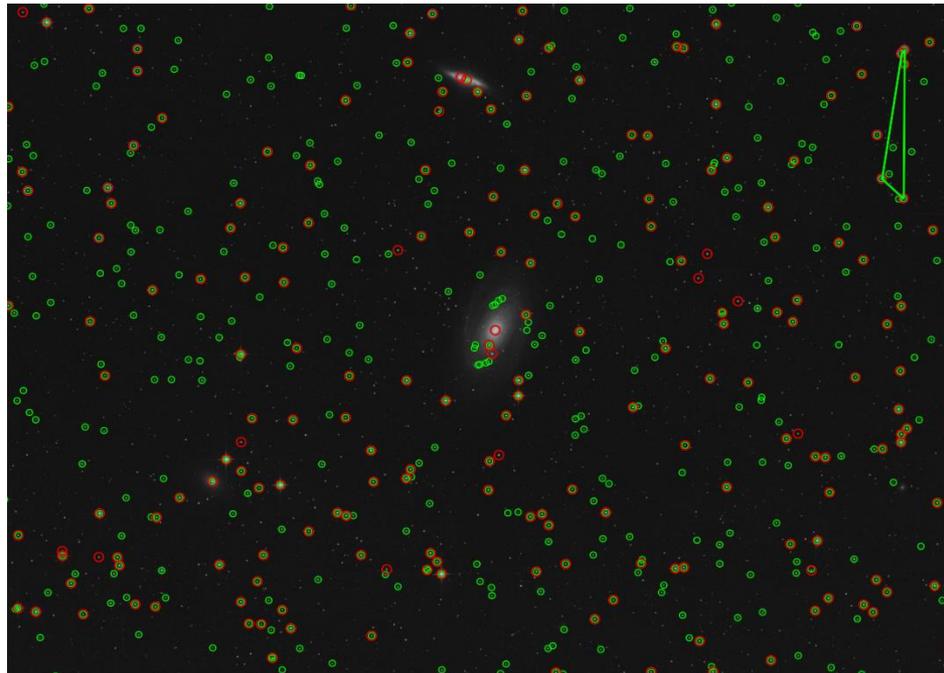


A shot of the Great Nebula, by Jerry Lodriguss (c.2006), from astropix.com
<http://astrometry.net/gallery.html>

Example



Example



A beautiful image of Bode's nebula (c.2007) by Peter Bressler, from starlightfriend.de
<http://astrometry.net/gallery.html>

Things to remember

- **Matching local invariant features**
 - Useful not only to provide matches for multi-view geometry, but also to find objects and scenes.
- **Bag of words** representation: quantize feature space to make discrete set of visual words
 - Summarize image by distribution of words
 - Index individual words
- **Inverted index**: pre-compute index to enable faster search at query time
- **Recognition of instances via alignment**: matching local features followed by spatial verification
 - Robust fitting : RANSAC, GHT

Discussion – Think-pair-share

- Find a person you don't know
- Discuss
 - strength,
 - weakness, and
 - potential extension
- Share with class

Image Categorization: Training phase

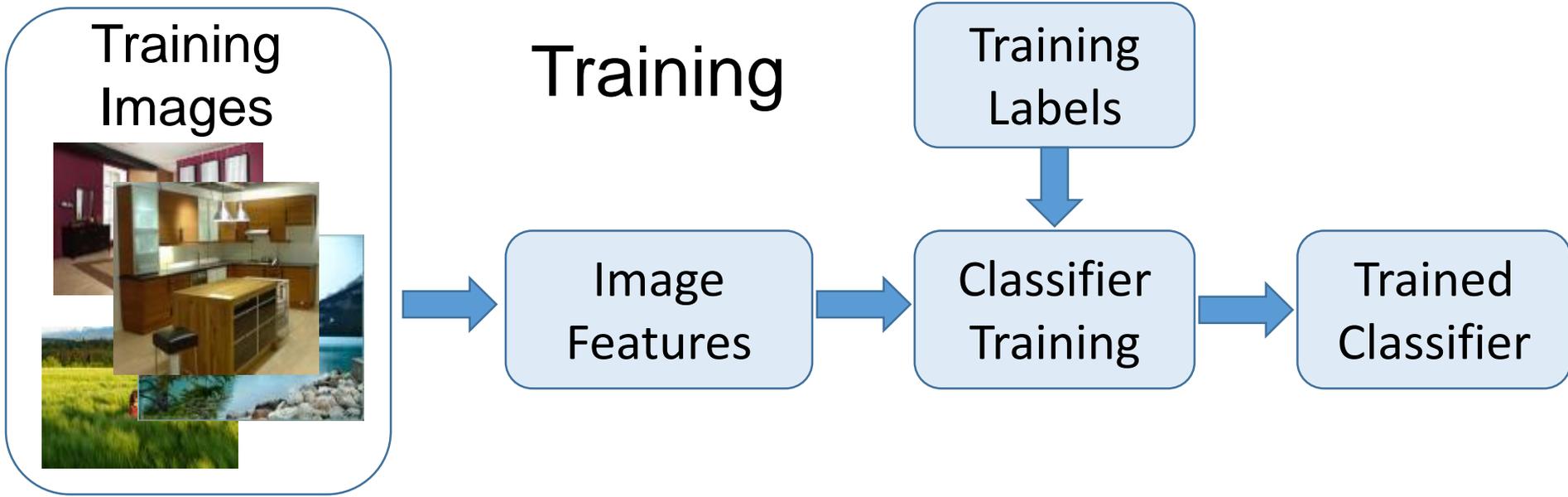
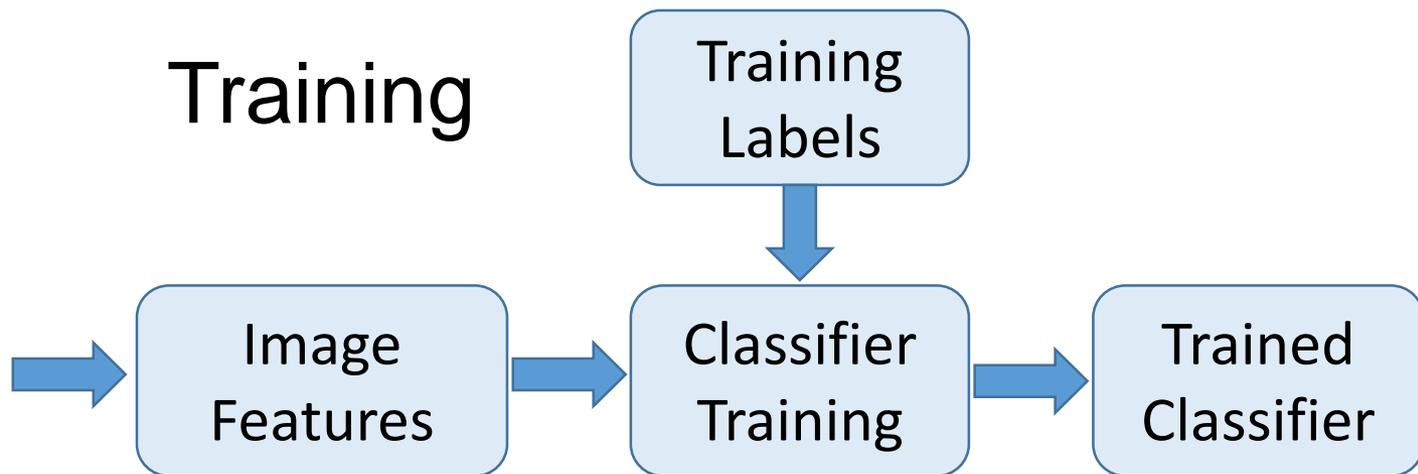


Image Categorization: Testing phase

Training Images



Training



Testing



Test Image

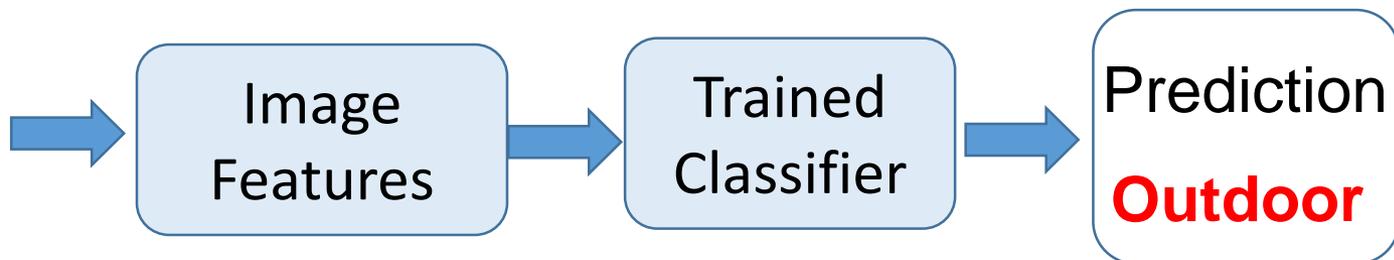


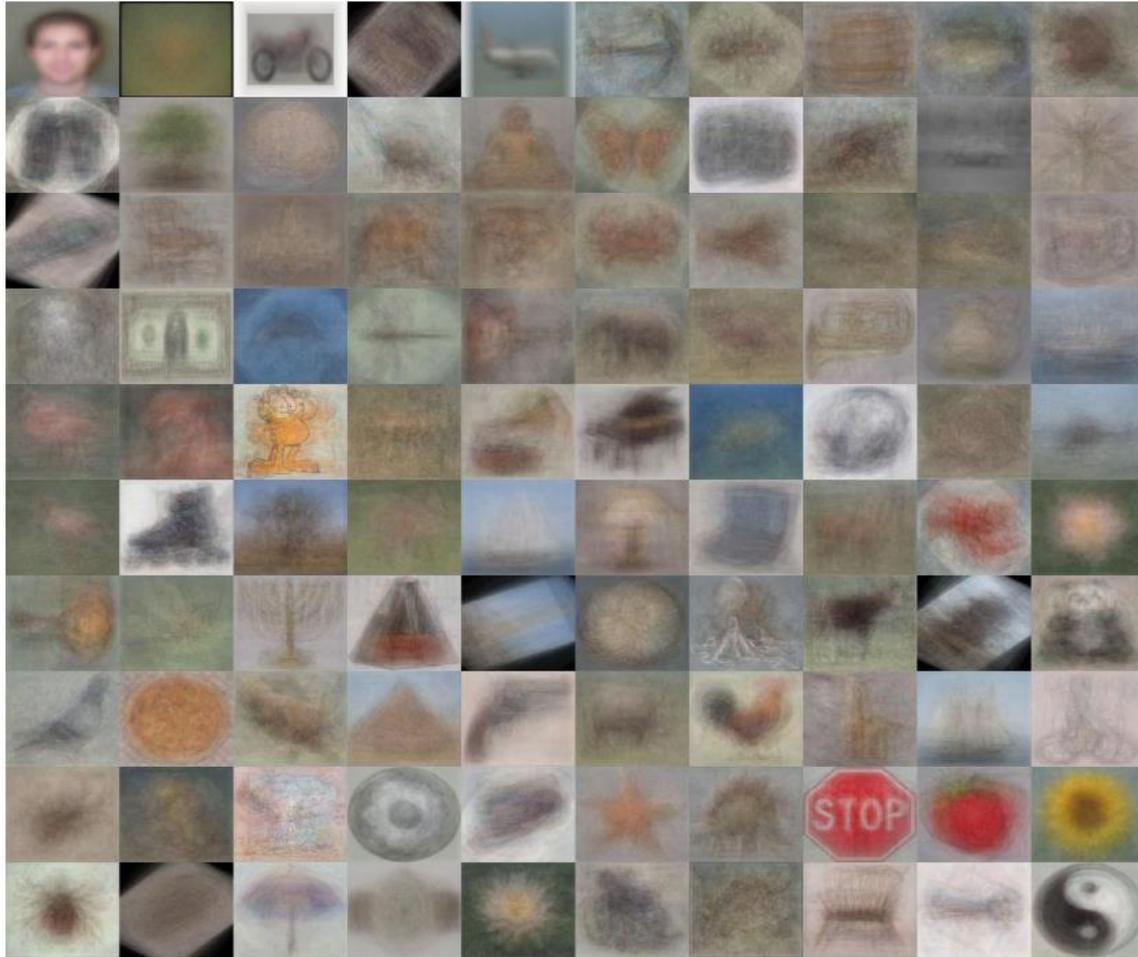
Image categorization

- Cat vs Dog



Image categorization

- Object recognition



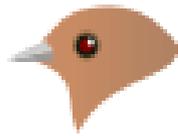
Caltech 101 Average Object Images

Image categorization

- Fine-grained recognition



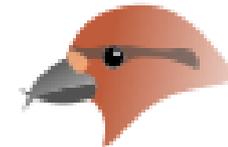
Generalist



Insect catching



Grain eating



Coniferous-seed eating



Nectar feeding



Chiseling



Dip netting



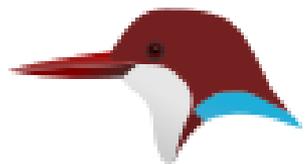
Surface skimming



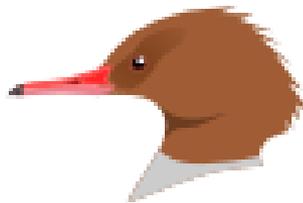
Scything



Probing



Aerial fishing



Pursuit fishing



Scavenging



Raptorial



Filter feeding

Image categorization

- Place recognition



Places Database [[Zhou et al. NIPS 2014](#)]

Image categorization

- Visual font recognition



Image categorization

- Dating historical photos



1940



1953



1966



1977

[[Palermo et al. ECCV 2012](#)]

Image categorization

- Image style recognition



HDR



Macro



Baroque



Roccoco



Vintage



Noir



Northern Renaissance



Cubism



Minimal



Hazy



Impressionism



Post-Impressionism



Long Exposure



Romantic



Abs. Expressionism



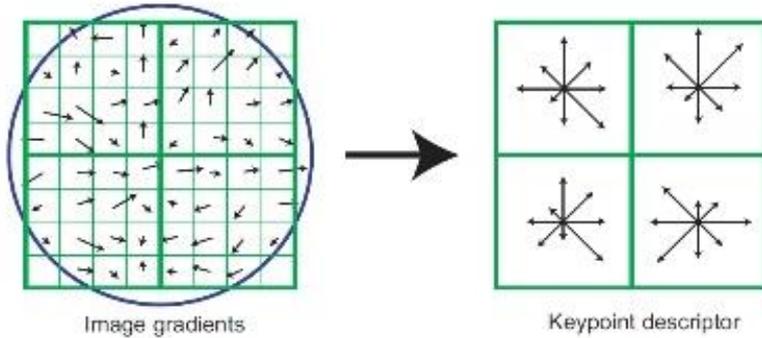
Color Field Painting

Flickr Style: 80K images covering 20 styles.

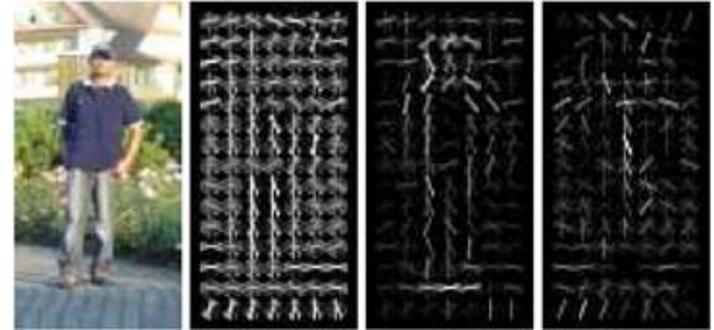
Wikipaintings: 85K images for 25 art genres.

[[Karayev et al. BMVC 2014](#)]

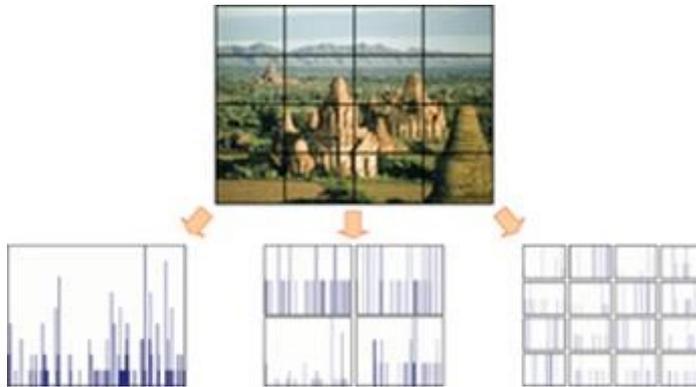
Features are the Keys



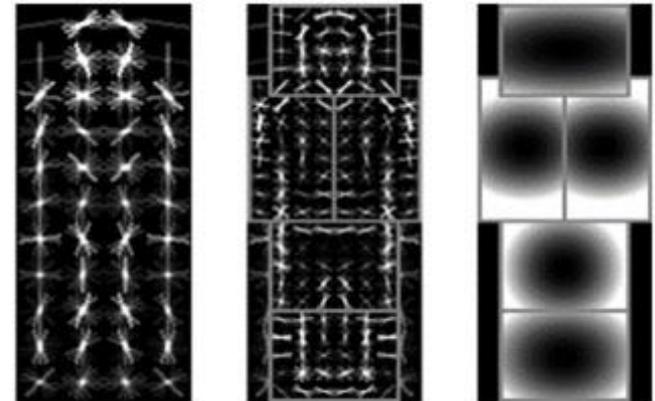
SIFT [[Loewe IJCV 04](#)]



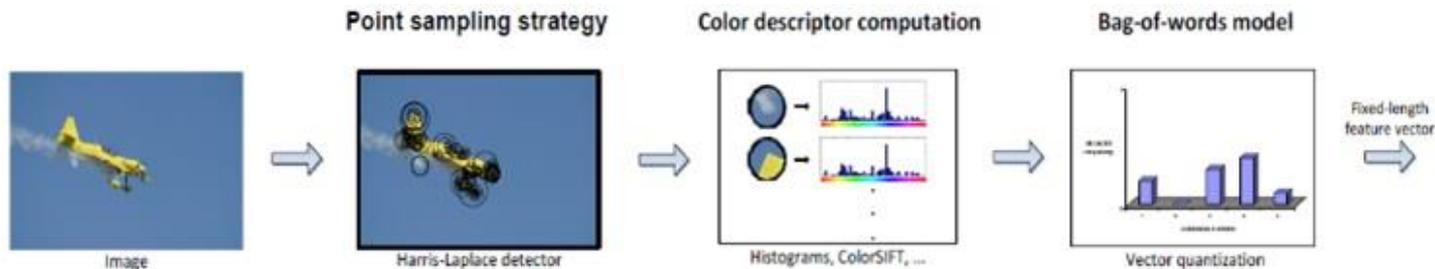
HOG [[Dalal and Triggs CVPR 05](#)]



SPM [[Lazebnik et al. CVPR 06](#)]



DPM [[Felzenszwalb et al. PAMI 10](#)]



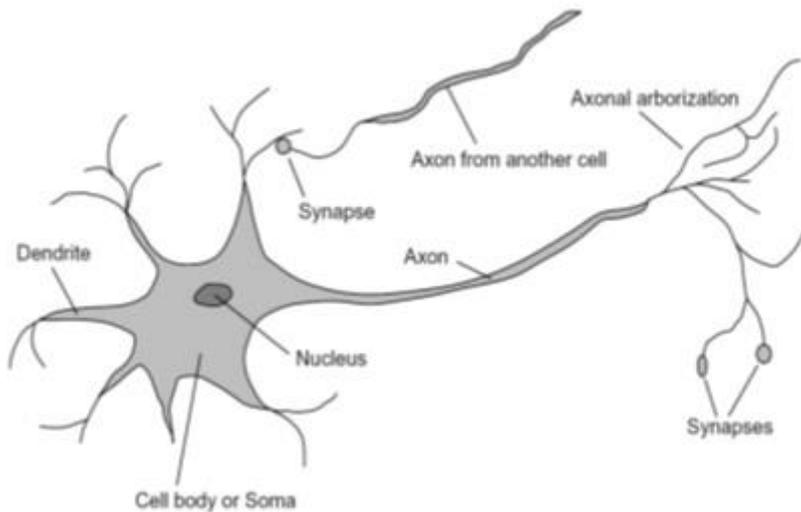
Color Descriptor [[Van De Sande et al. PAMI 10](#)]

Learning a Hierarchy of Feature Extractors

- Each layer of hierarchy extracts features from output of previous layer
- All the way from pixels \rightarrow classifier
- Layers have the (nearly) same structure

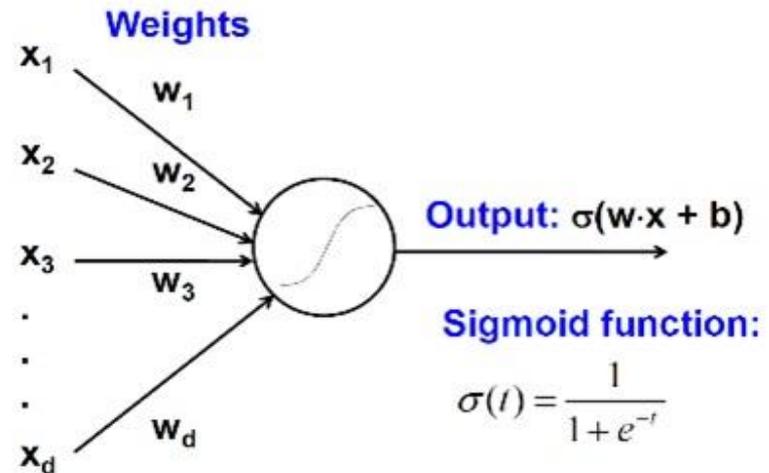


Biological neuron and Perceptrons

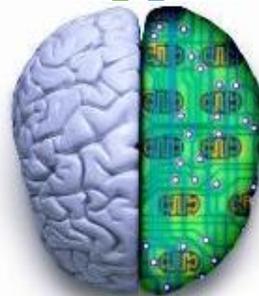


A biological neuron

Input



An artificial neuron (Perceptron)
- a linear classifier



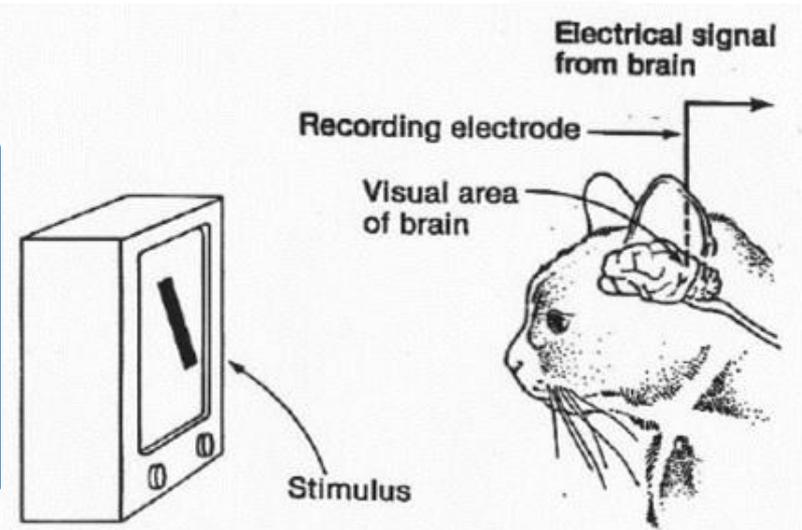
Simple, Complex and Hypercomplex cells



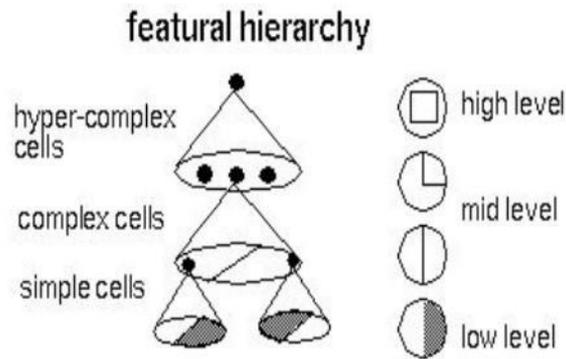
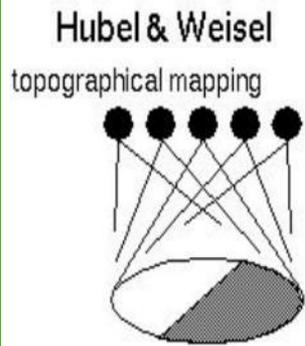
David H. Hubel and Torsten Wiesel

Suggested a **hierarchy of feature detectors** in the visual cortex, with higher level features responding to patterns of activation in lower level cells, and propagating activation upwards to still higher level cells.

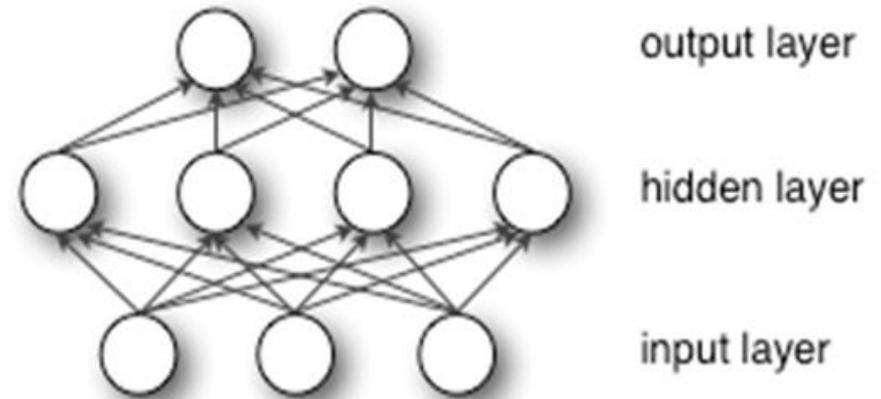
David Hubel's [Eye, Brain, and Vision](#)



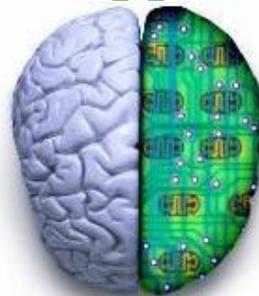
Hubel/Wiesel Architecture and Multi-layer Neural Network



Hubel and Wiesel's architecture



Multi-layer Neural Network
- *A non-linear classifier*

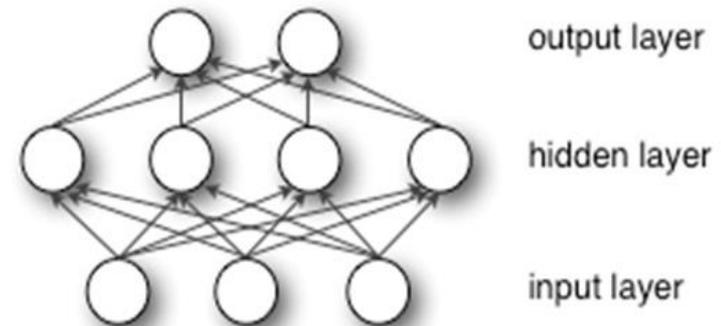


Multi-layer Neural Network

- A non-linear classifier
- **Training:** find network weights \mathbf{w} to minimize the error between true training labels y_i and estimated labels $f_{\mathbf{w}}(\mathbf{x}_i)$

$$E(\mathbf{w}) = \sum_{i=1}^N (y_i - f_{\mathbf{w}}(\mathbf{x}_i))^2$$

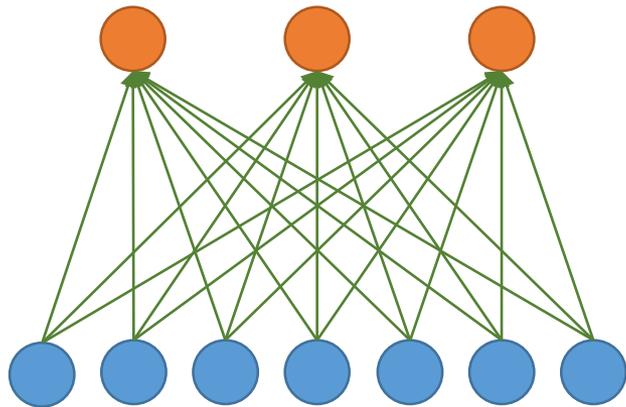
- Minimization can be done by gradient descent provided f is differentiable
- This training method is called [back-propagation](#)



Convolutional Neural Networks

- Also known as CNN, ConvNet, DCN
- CNN = a multi-layer neural network with
 1. Local connectivity
 2. Weight sharing

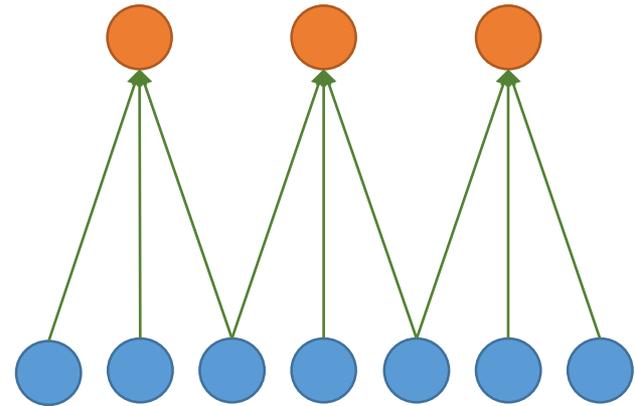
CNN: Local Connectivity



Global connectivity

Hidden layer

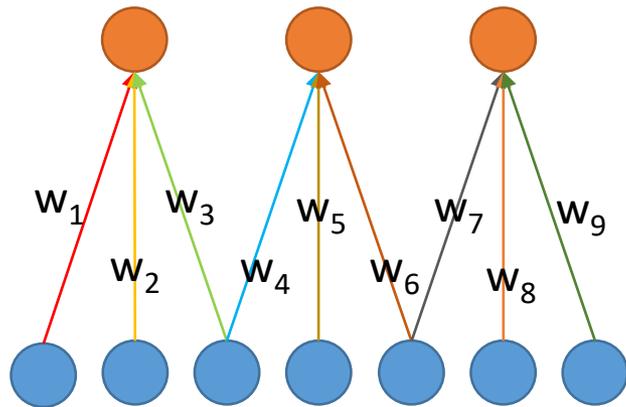
Input layer



Local connectivity

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

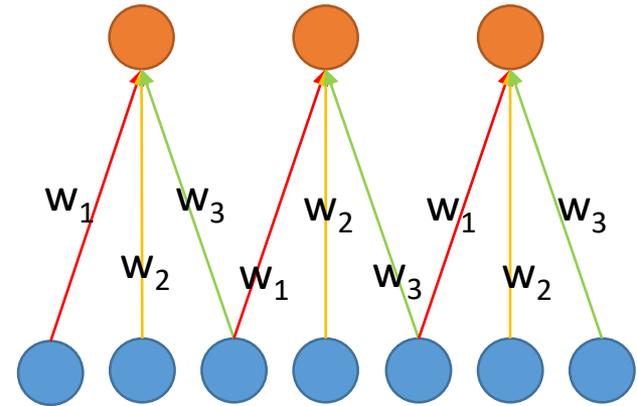
CNN: Weight Sharing



Hidden layer

Input layer

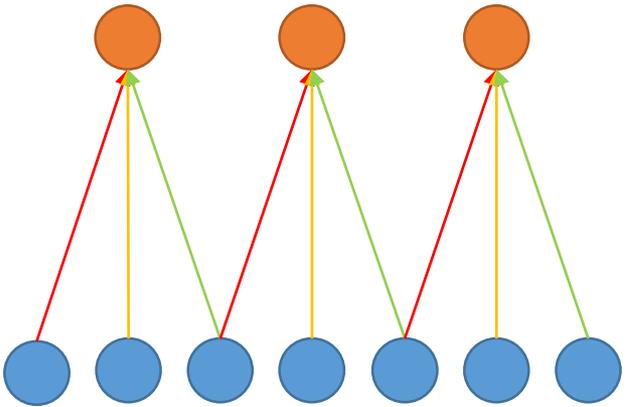
Without weight sharing



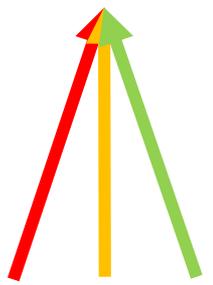
With weight sharing

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

CNN with multiple input channels



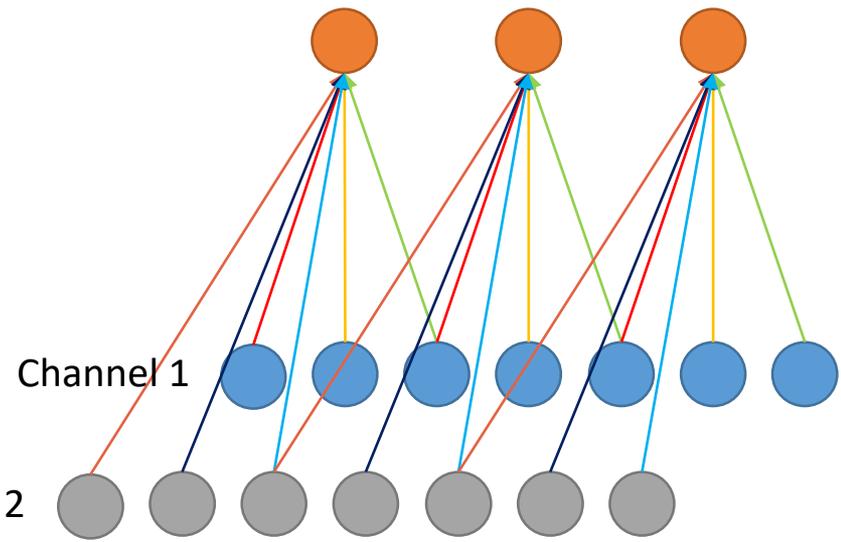
Single input channel



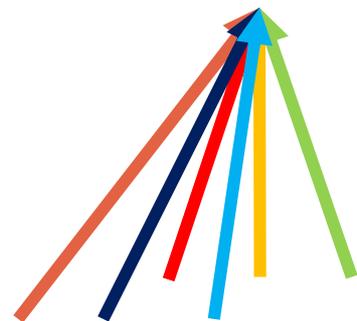
Filter weights

Hidden layer

Input layer

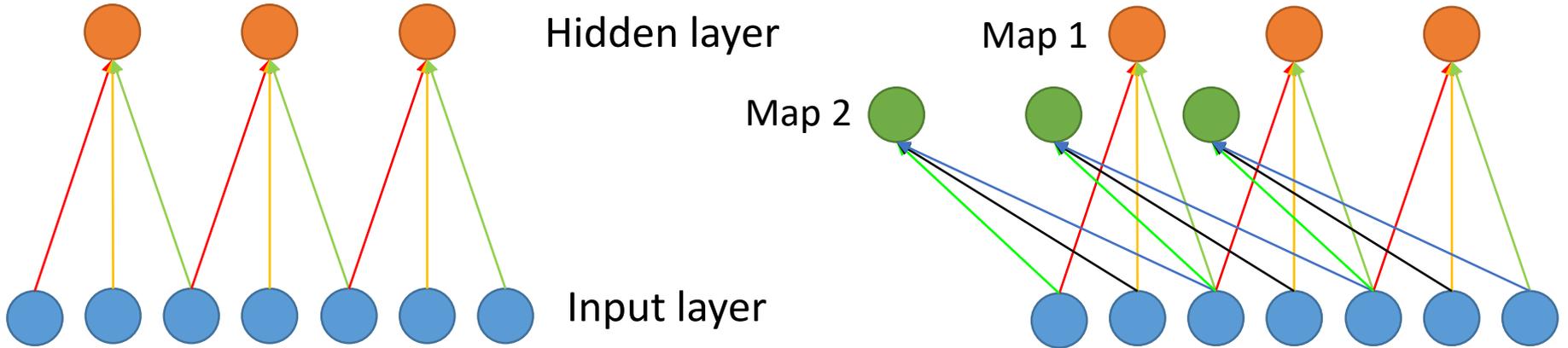


Multiple input channels

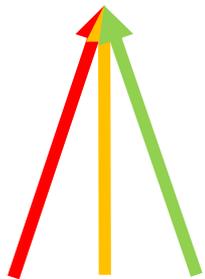


Filter weights

CNN with multiple output maps

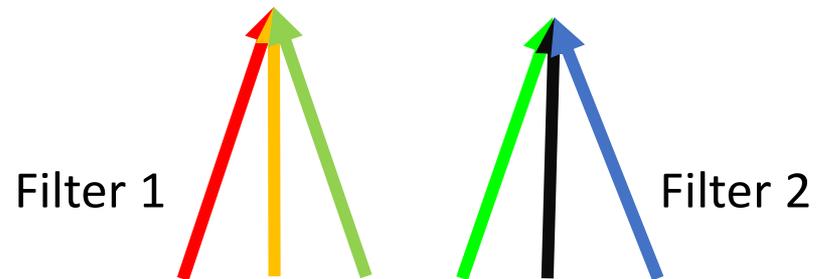


Single output map



Filter weights

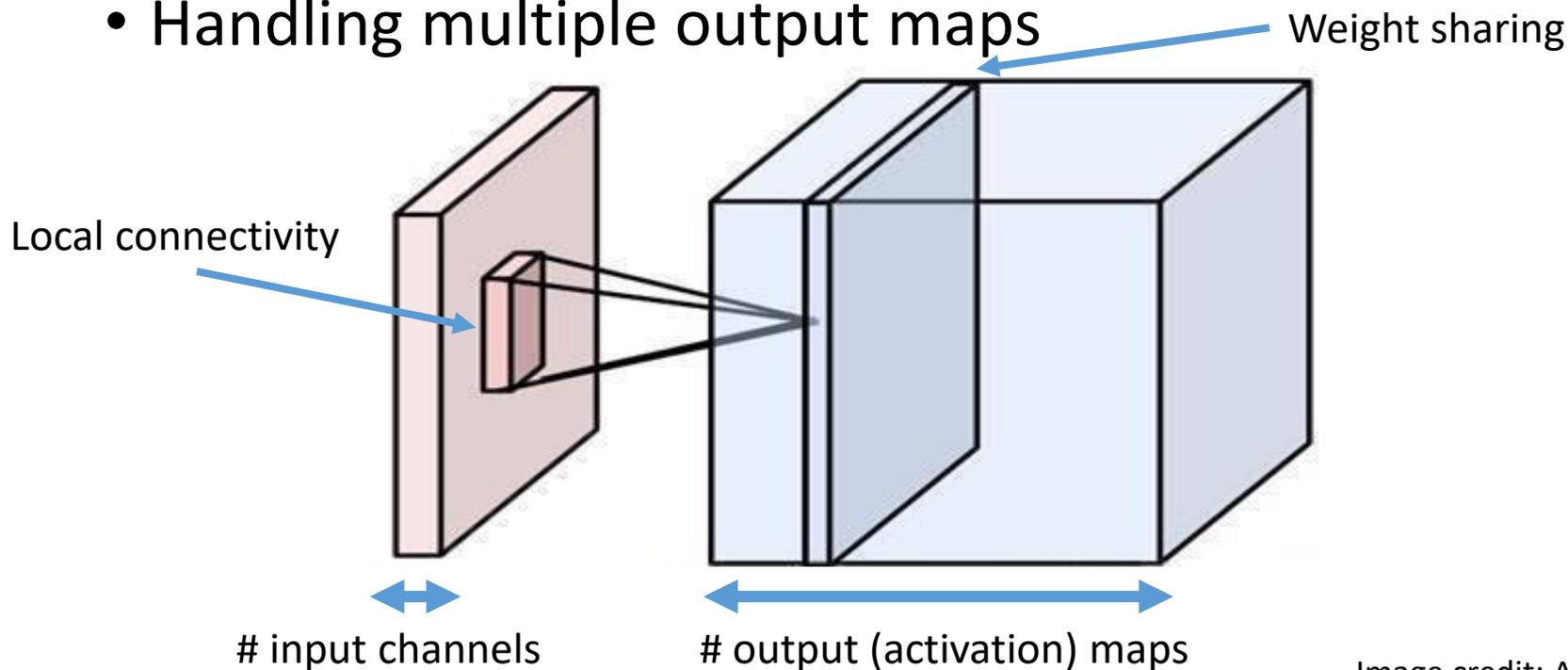
Multiple output maps



Filter weights

Putting them together

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



Neocognitron [[Fukushima, Biological Cybernetics 1980](#)]

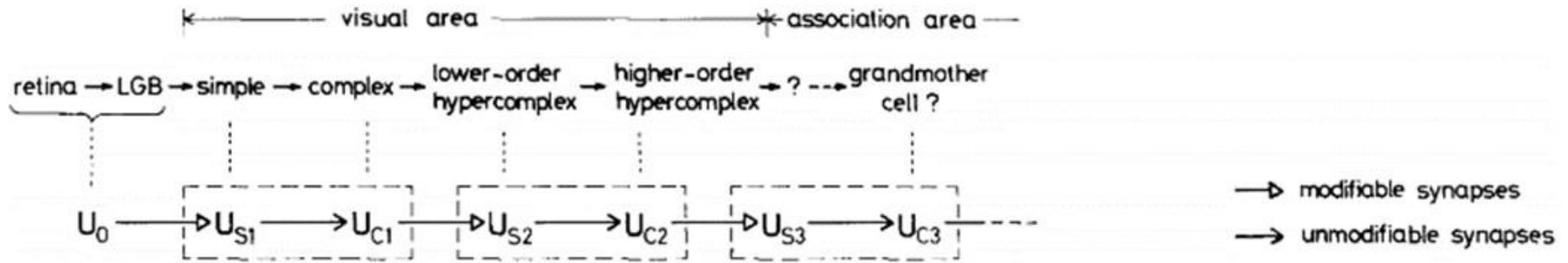
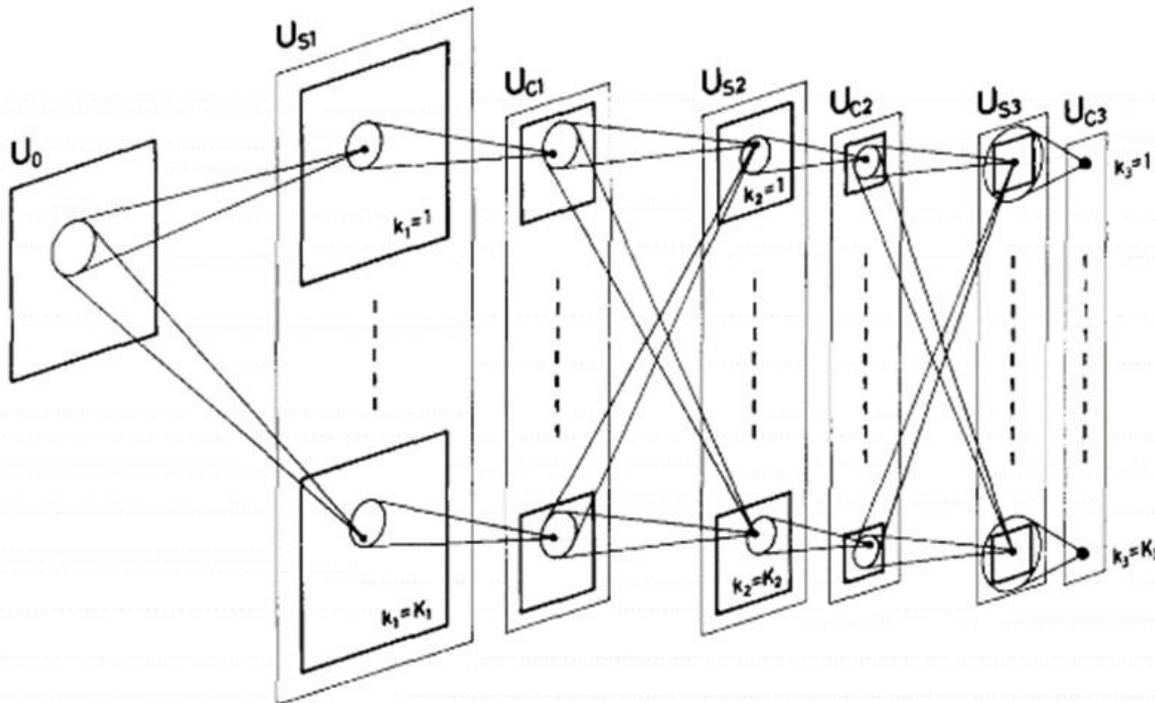


Fig. 1. Correspondence between the hierarchy model by Hubel and Wiesel, and the neural network of the neocognitron

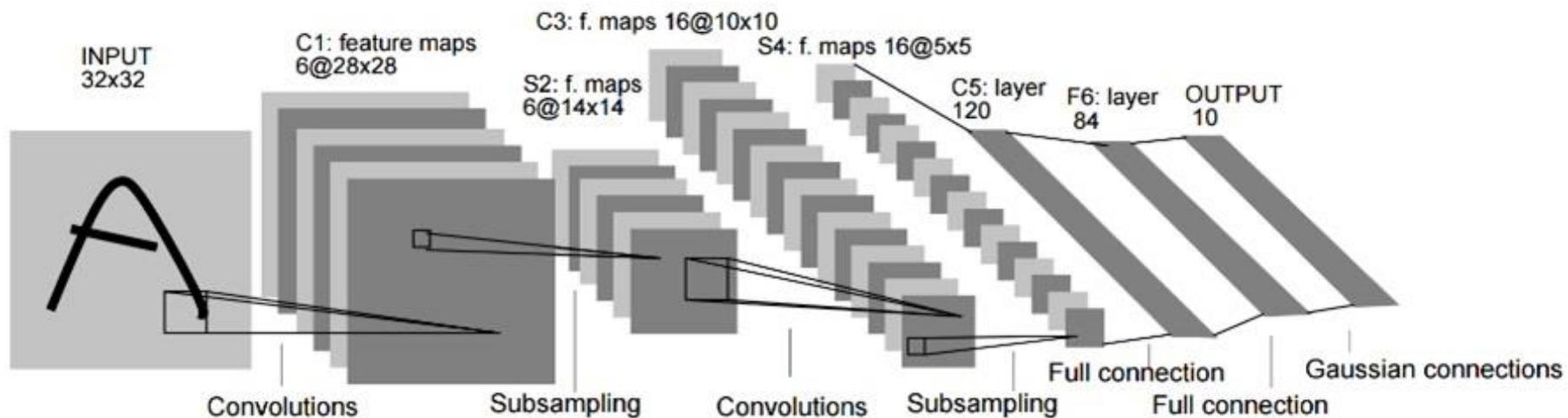


Deformation-Resistant
Recognition

S-cells: (simple)
- extract local features

C-cells: (complex)
- allow for positional errors

LeNet [LeCun et al. 1998]



Gradient-based learning applied to document recognition [[LeCun, Bottou, Bengio, Haffner 1998](#)]

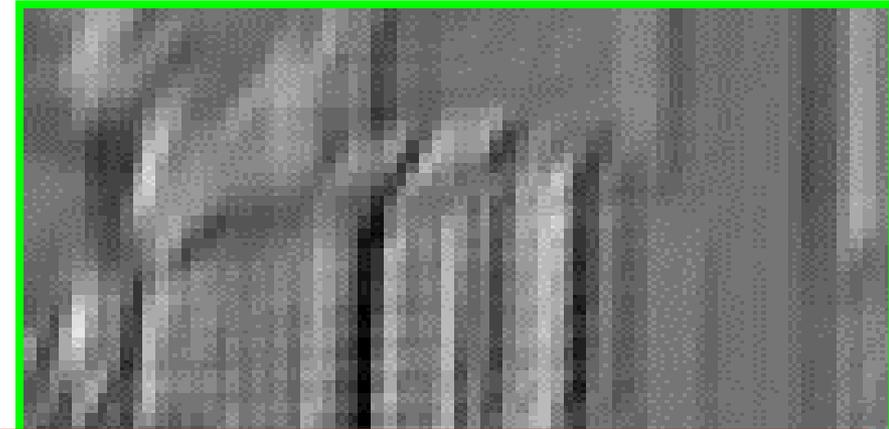
LeNet-1 from 1993

What is a Convolution?

- Weighted moving sum



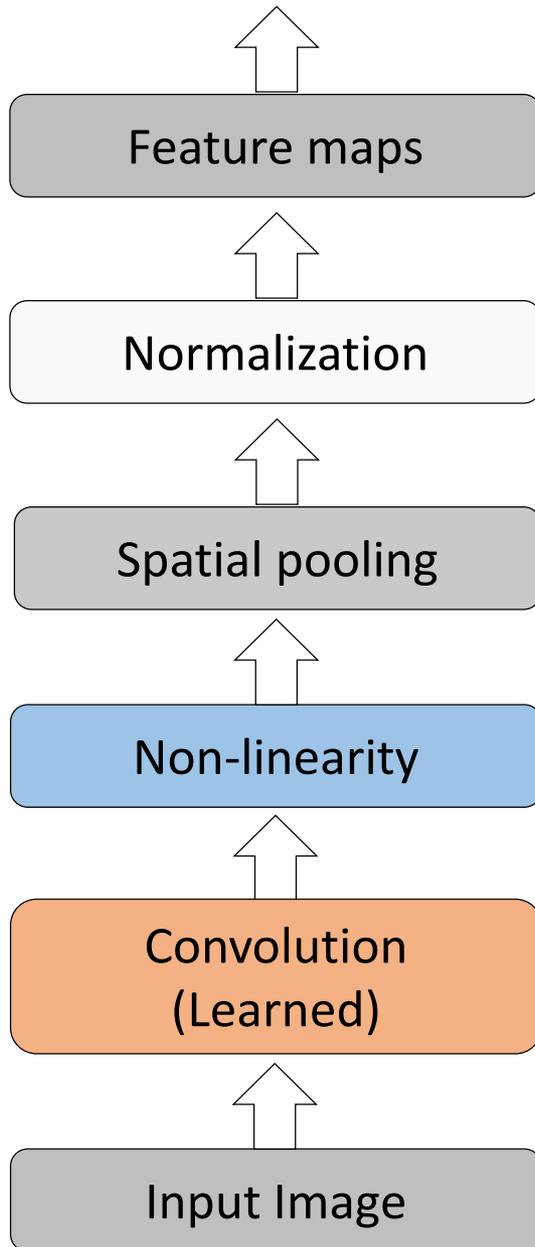
Input



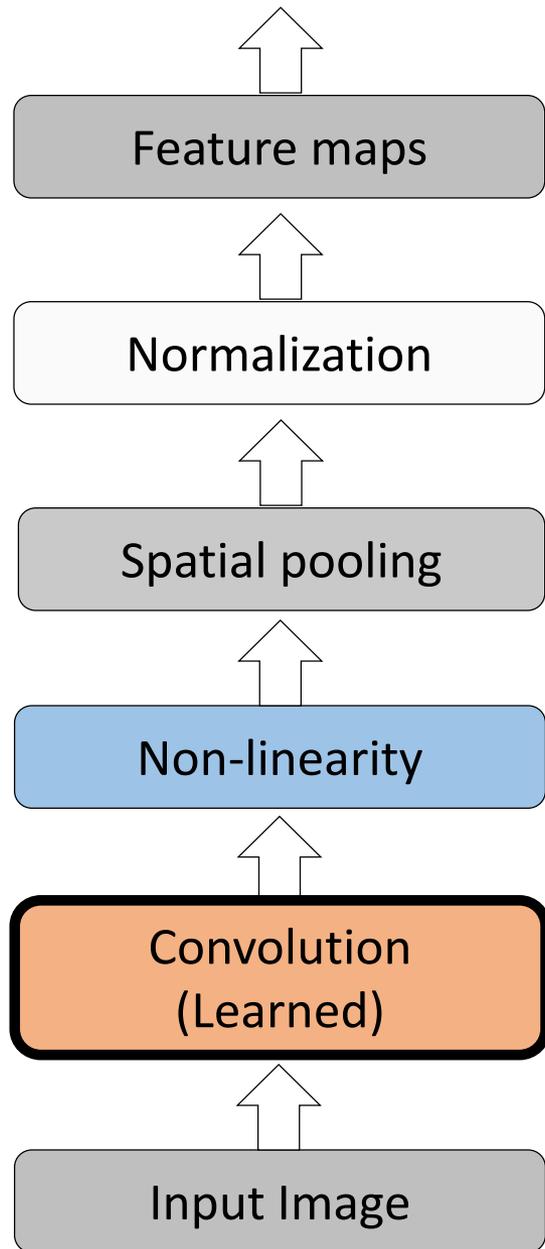
Feature Activation Map

slide credit: S. Lazebnik

Convolutional Neural Networks



Convolutional Neural Networks

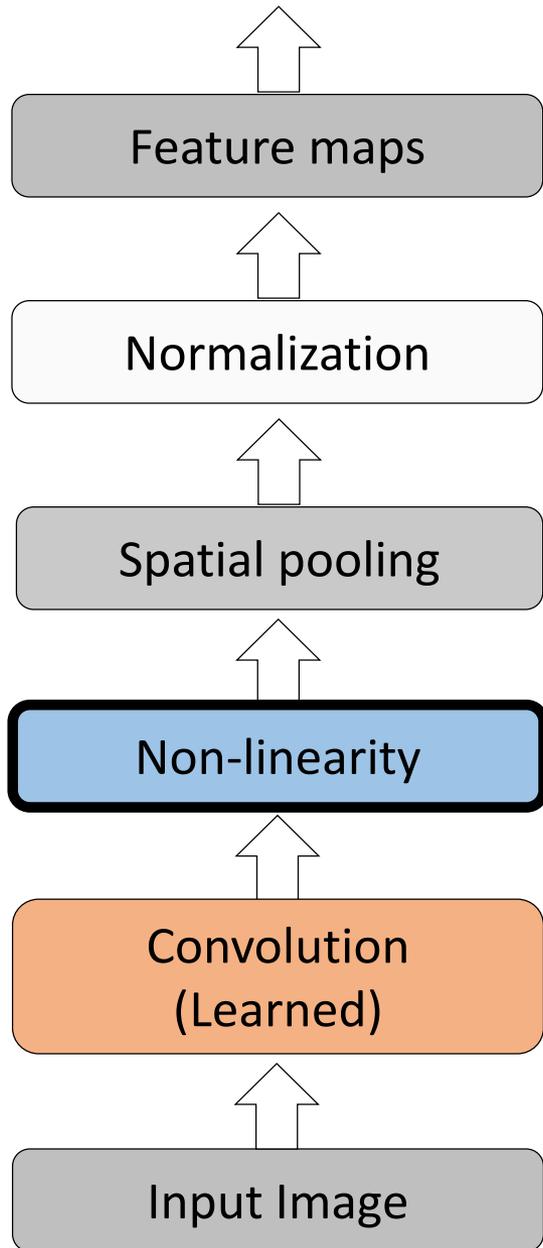


Input

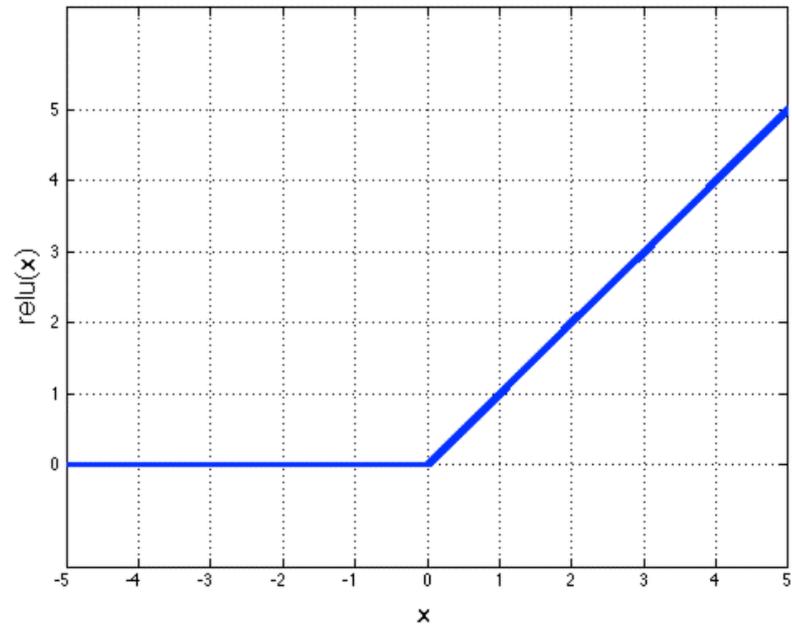


Feature Map

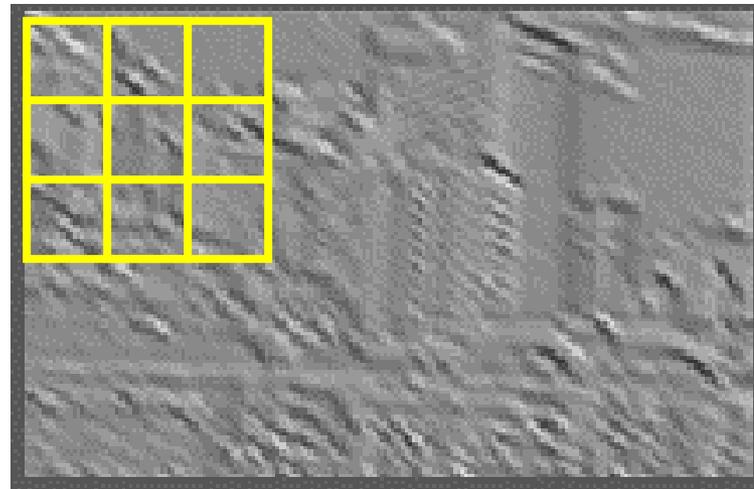
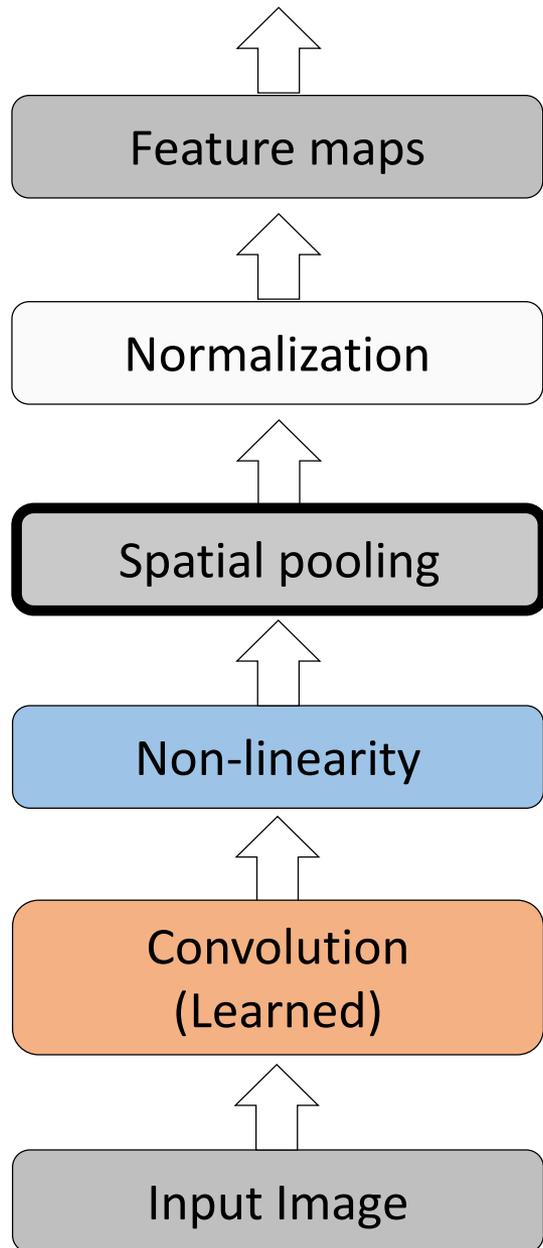
Convolutional Neural Networks



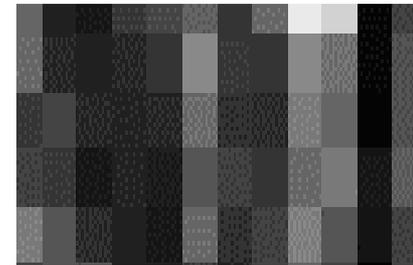
Rectified Linear Unit (ReLU)



Convolutional Neural Networks



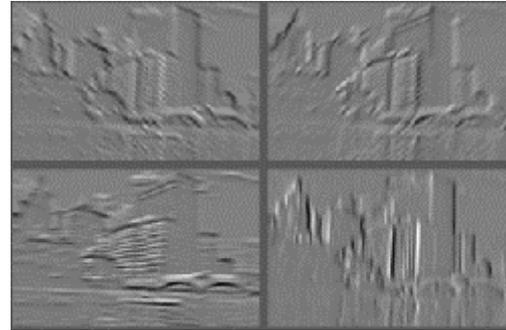
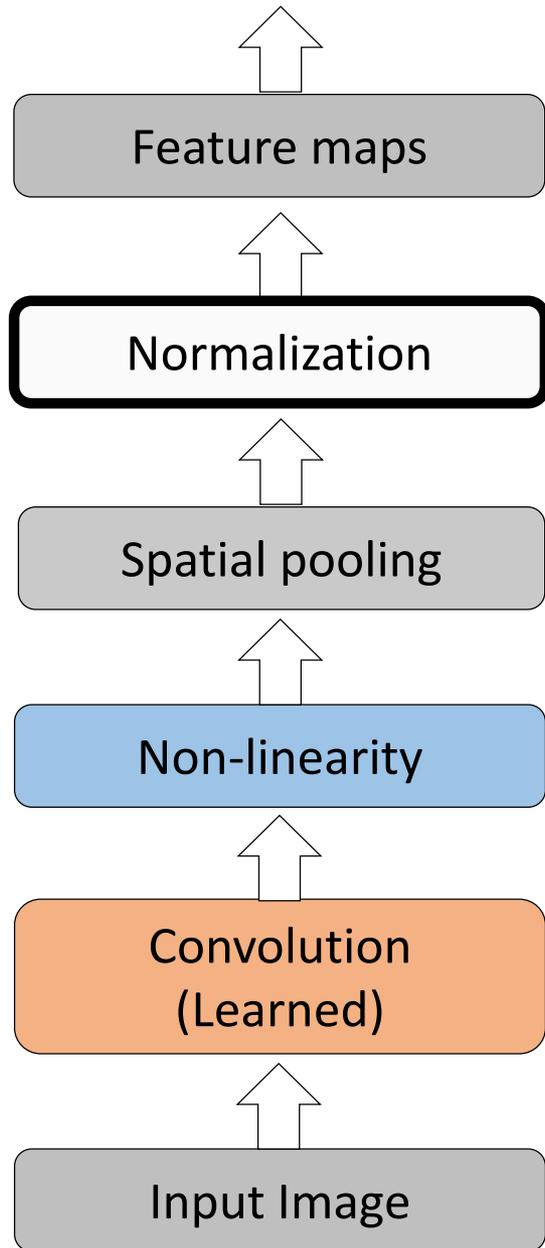
Max pooling



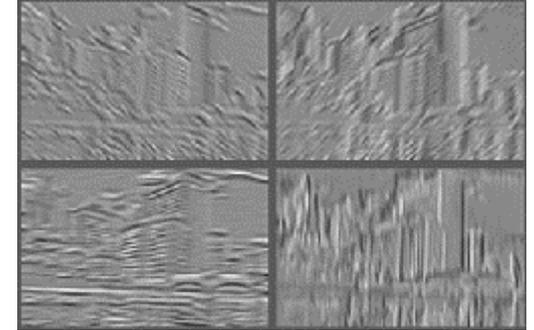
Max-pooling: a non-linear down-sampling

Provide *translation invariance*

Convolutional Neural Networks

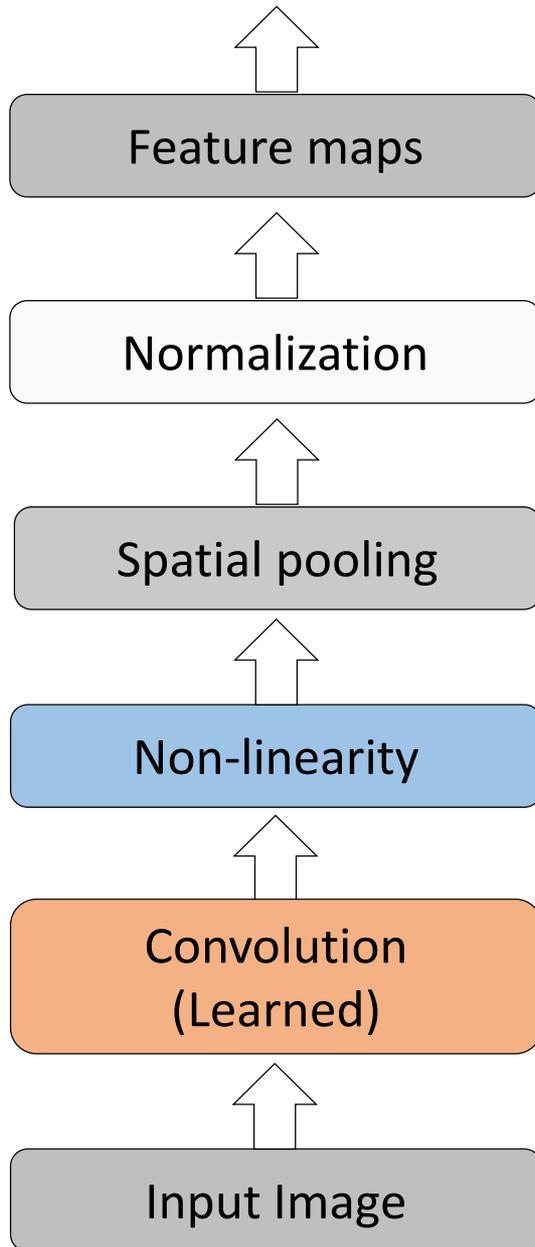


Feature Maps



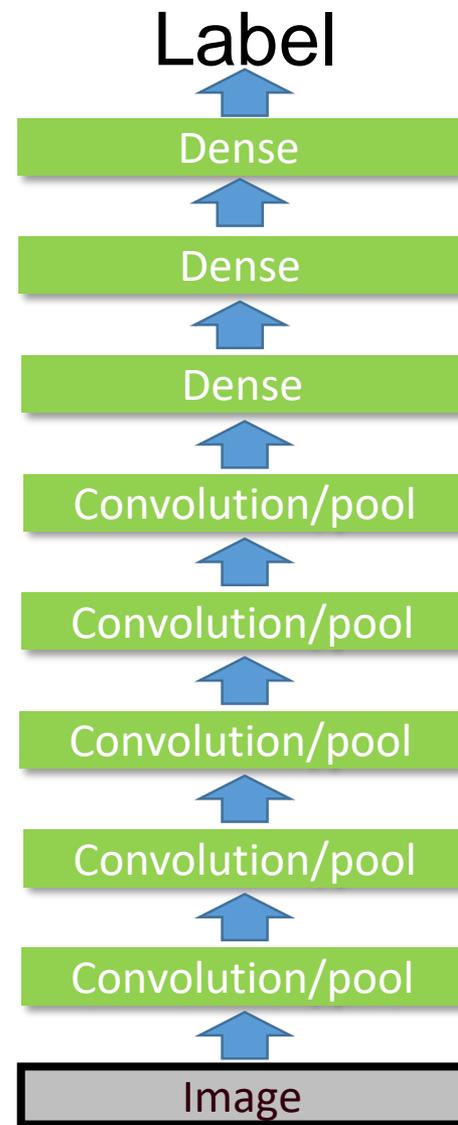
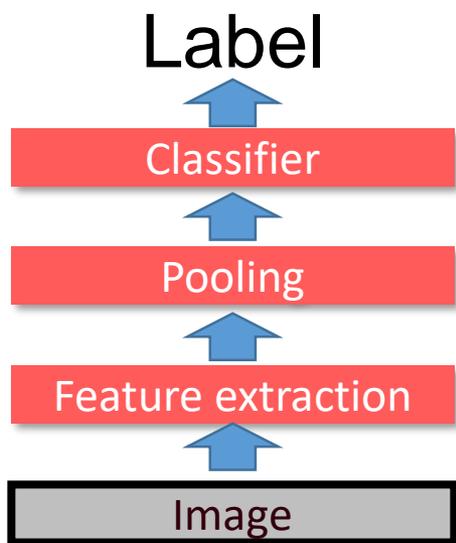
Feature Maps
After Contrast
Normalization

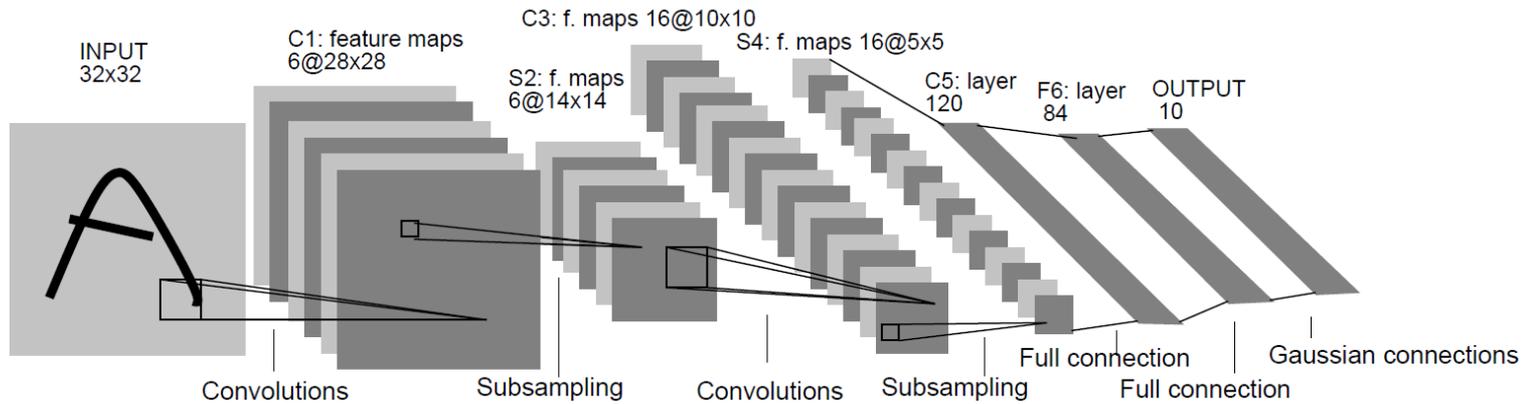
Convolutional Neural Networks



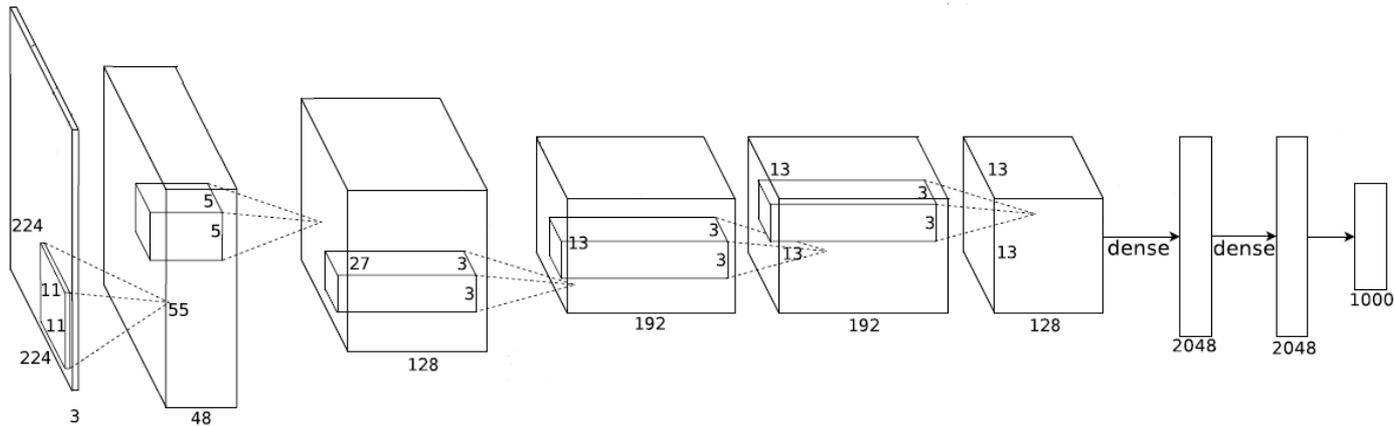
Engineered vs. learned features

Convolutional filters are trained in a supervised manner by back-propagating classification error

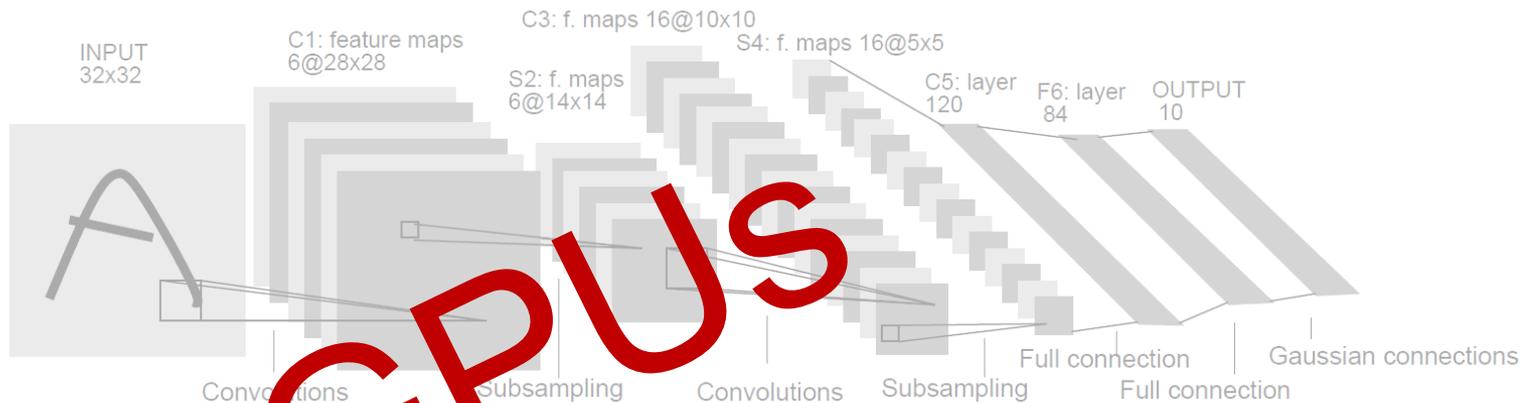




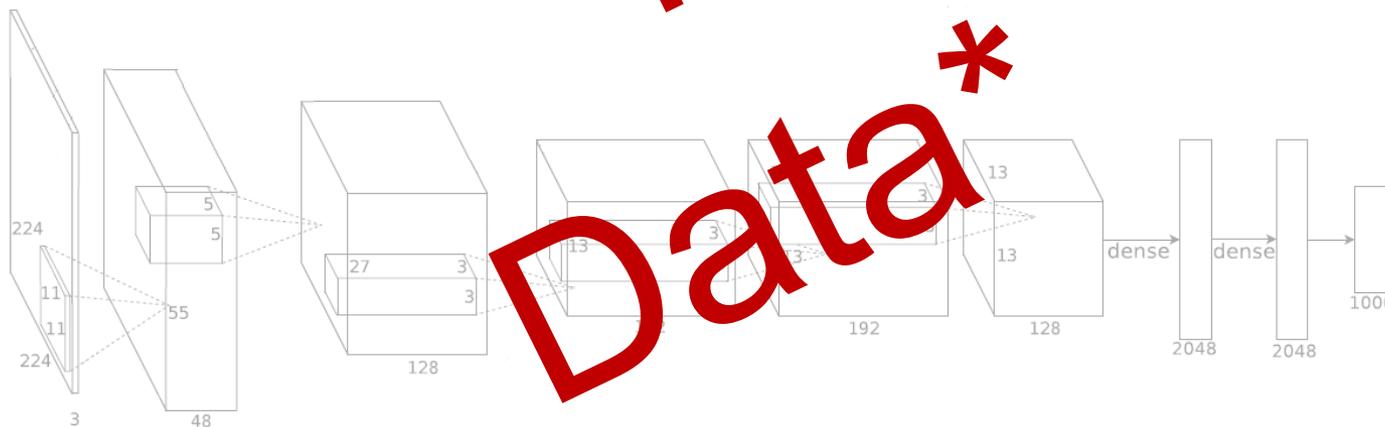
Gradient-Based Learning Applied to Document Recognition, LeCun, Bottou, Bengio and Haffner, Proc. of the IEEE, **1998**



Imagenet Classification with Deep Convolutional Neural Networks, Krizhevsky, Sutskever, and Hinton, NIPS **2012**



Gradient-Based Learning Applied to Document Recognition, LeCun, Bottou, Bengio and Haffner, Proc. of the IEEE, 1998



Imagenet Classification
Networks, Krizhevsky et al.

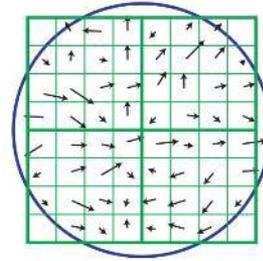
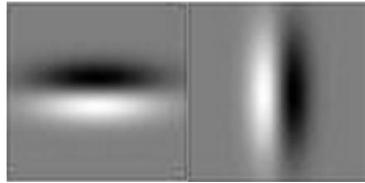
* Rectified activations and dropout

SIFT Descriptor

Image
Pixels

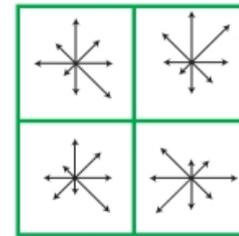
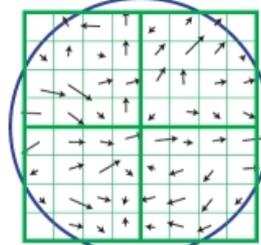


Apply gradient
filters

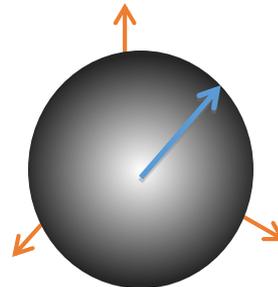


Lowé [IJCV 2004]

Spatial pool
(Sum)



Normalize to unit
length



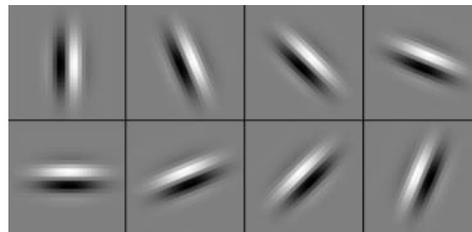
Feature
Vector

SIFT Descriptor

Image
Pixels

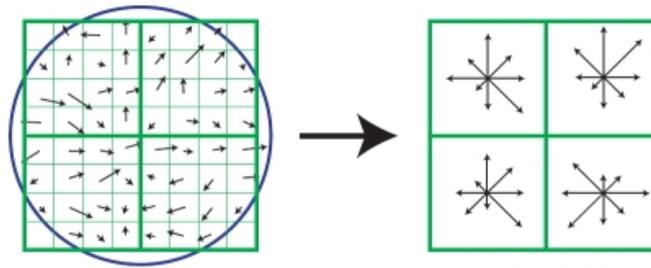


Apply
oriented filters

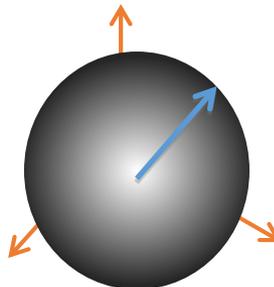


Lowé [IJCV 2004]

Spatial pool
(Sum)



Normalize to unit
length



Feature
Vector



slide credit: R. Fergus

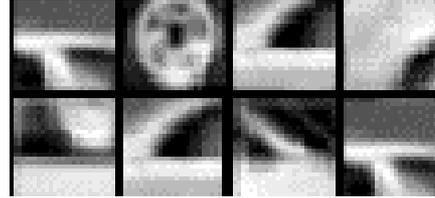
Spatial Pyramid Matching

Lazebnik,
Schmid,
Ponce
[CVPR 2006]

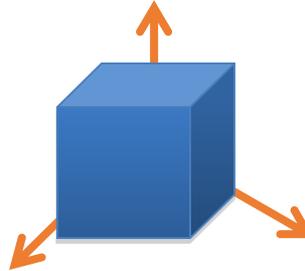
SIFT
Features



Filter with
Visual Words



Max



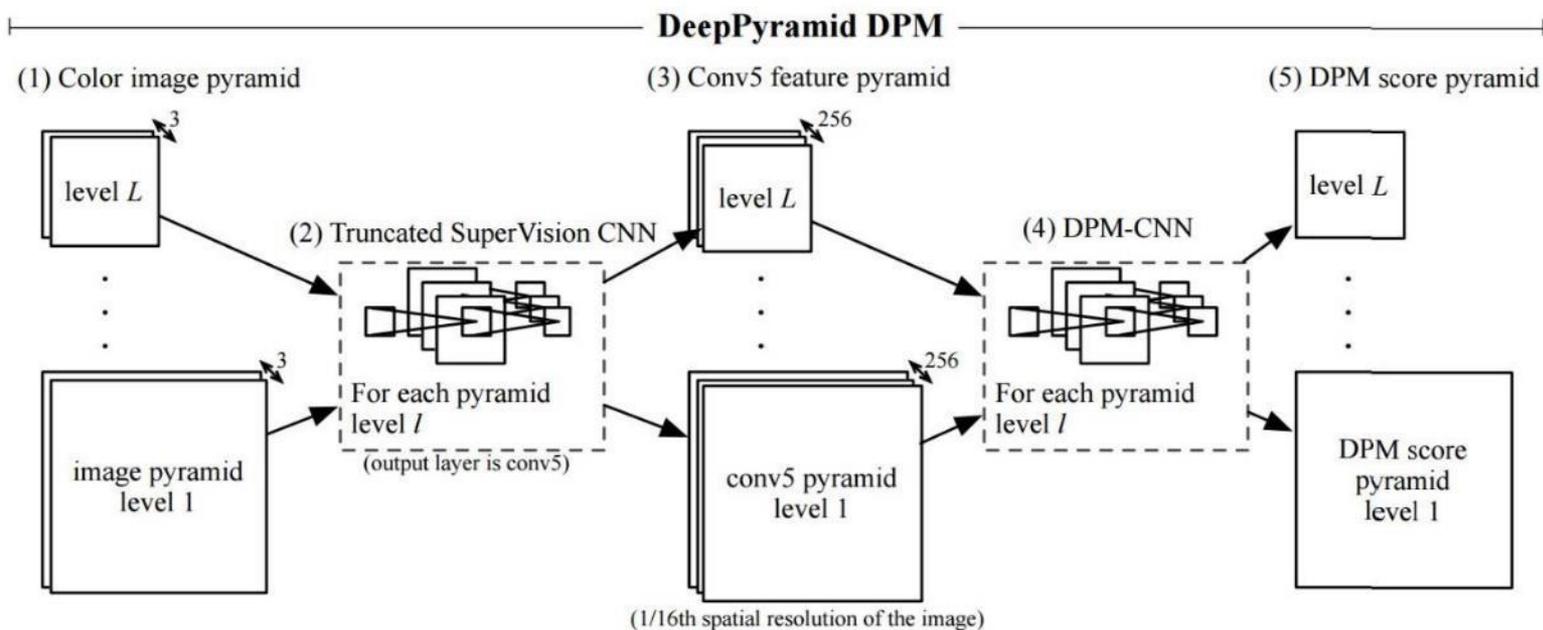
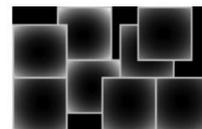
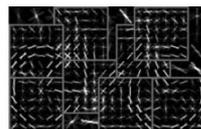
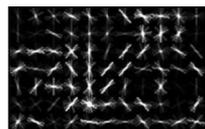
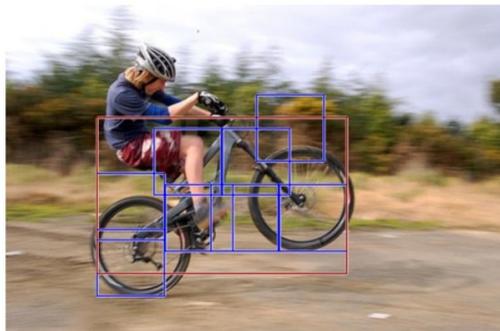
Multi-scale
spatial pool
(Sum)



Classifier

slide credit: R. Fergus

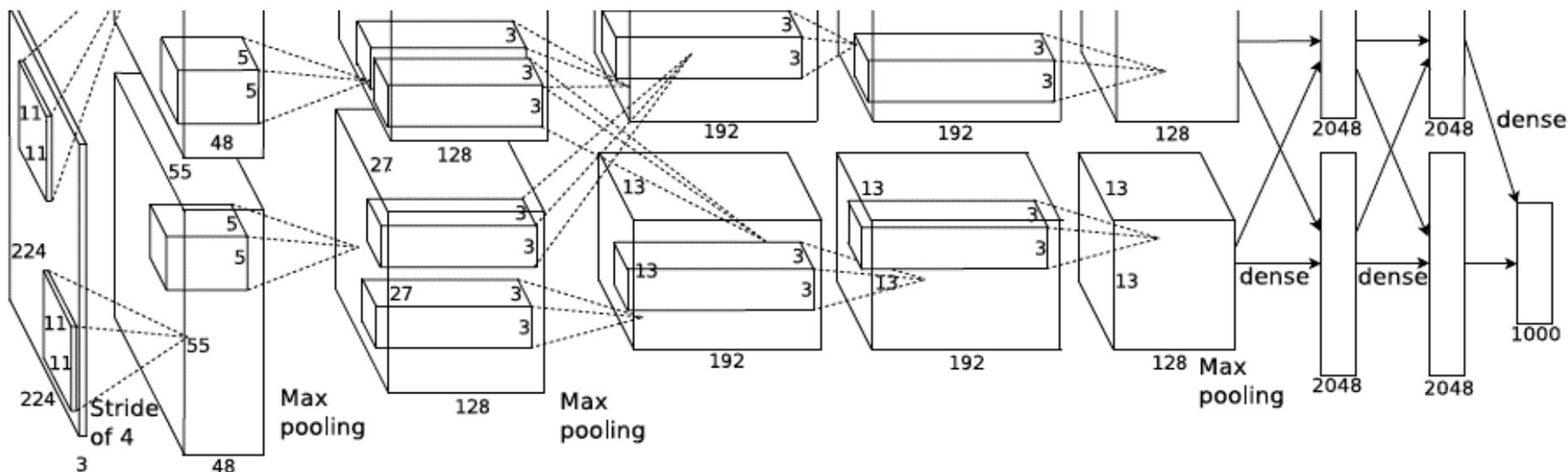
Deformable Part Model



Deformable Part Models are Convolutional Neural Networks [[Girshick et al. CVPR 15](#)]

AlexNet

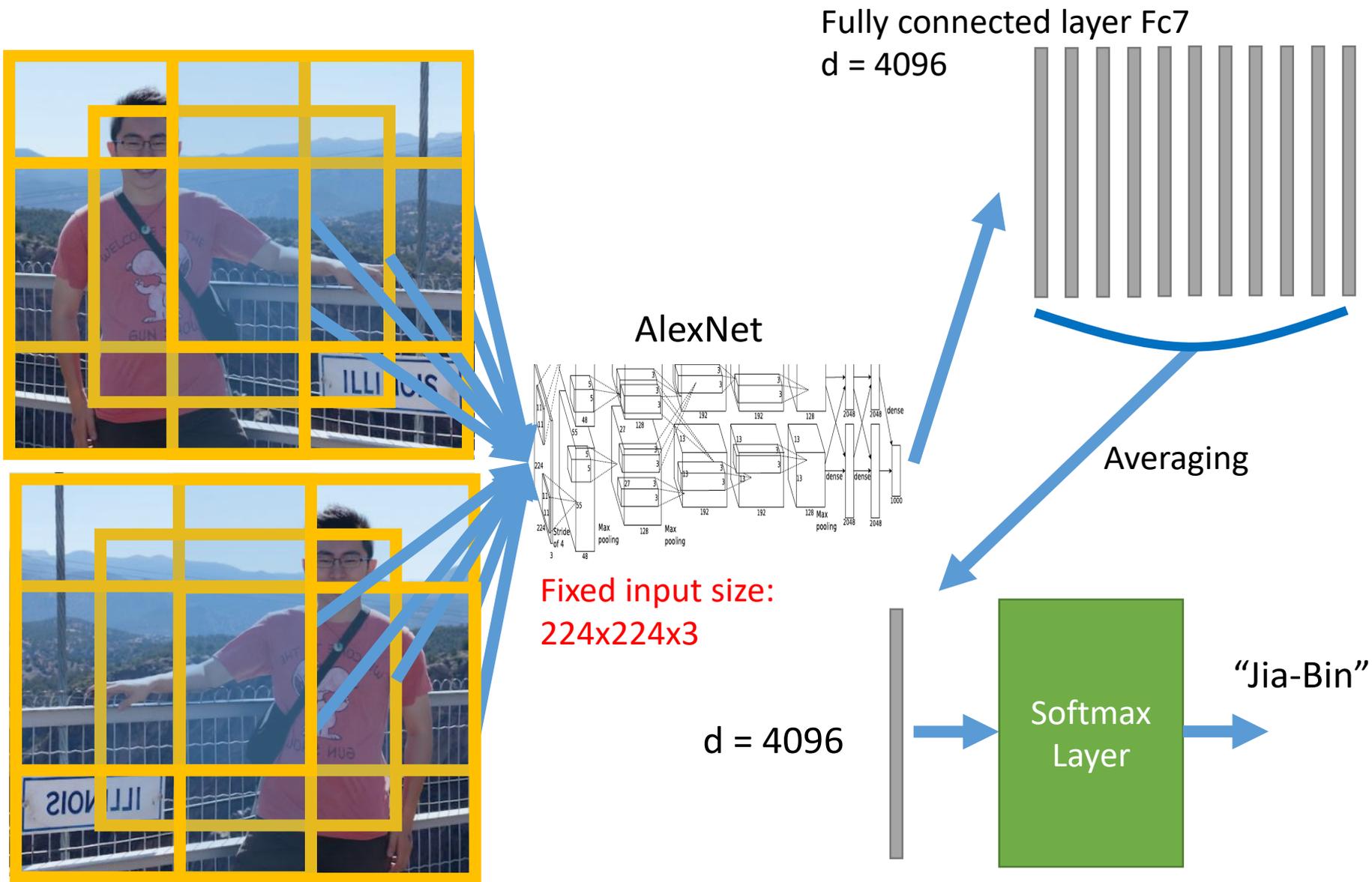
- Similar framework to LeCun'98 but:
 - Bigger model (7 hidden layers, 650,000 units, 60,000,000 params)
 - More data (10^6 vs. 10^3 images)
 - GPU implementation (50x speedup over CPU)
 - Trained on two GPUs for a week



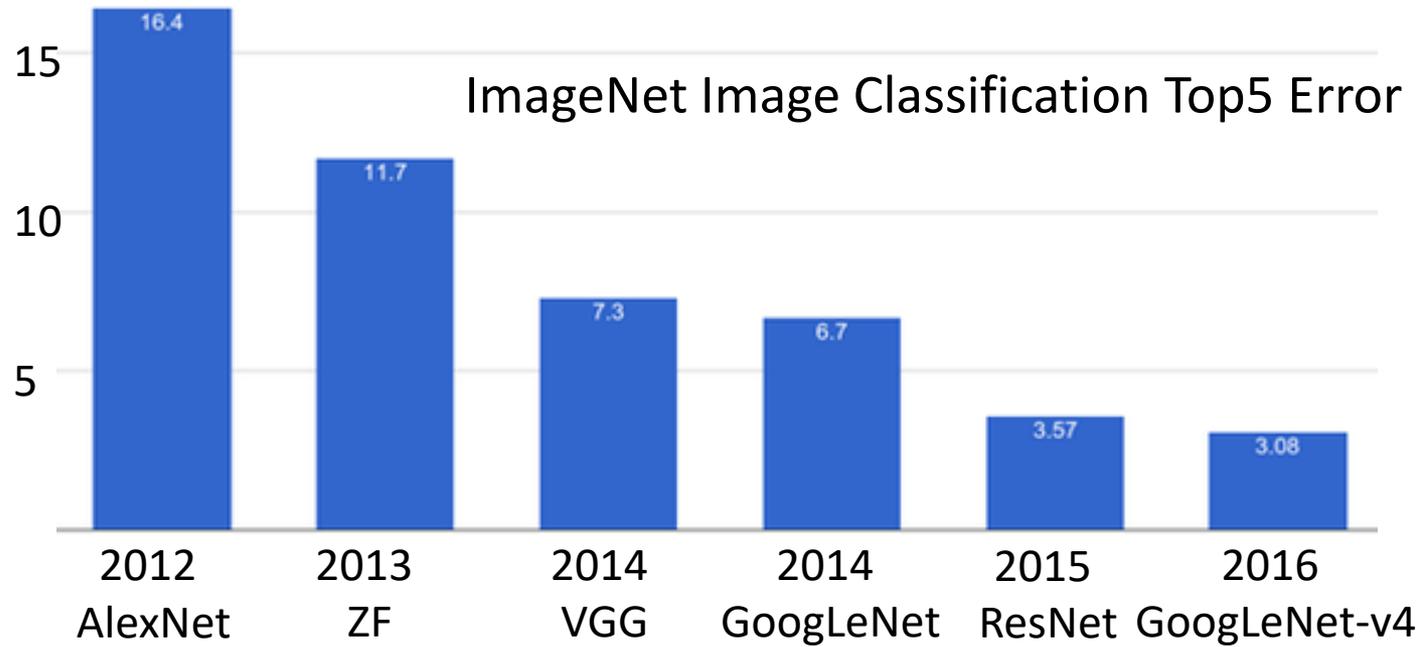
A. Krizhevsky, I. Sutskever, and G. Hinton,

[ImageNet Classification with Deep Convolutional Neural Networks](#), NIPS 2012

Using CNN for Image Classification



Progress on ImageNet



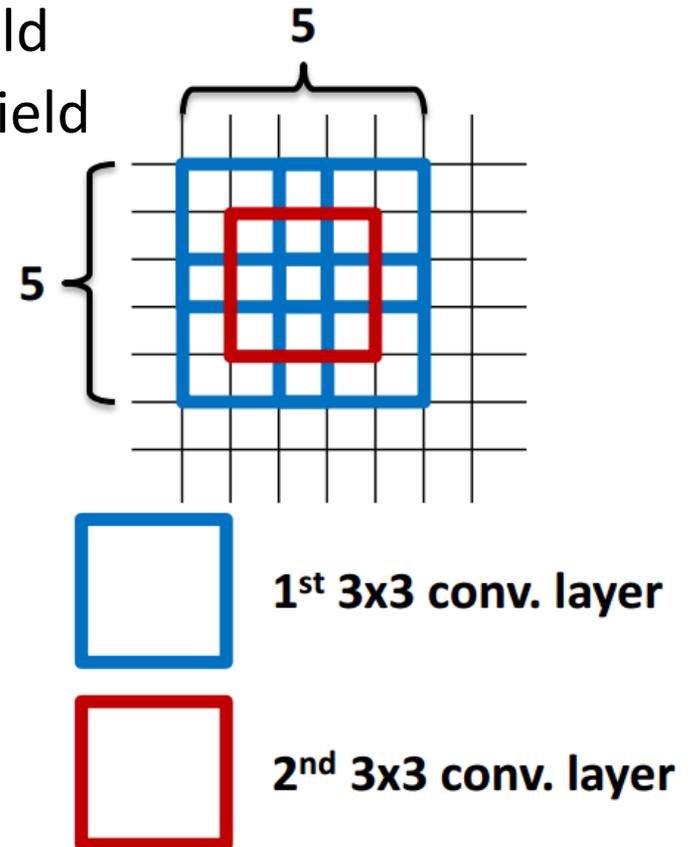
VGG-Net

- The deeper, the better
- Key design choices:
 - 3x3 conv. Kernels
 - very small
 - conv. stride 1
 - no loss of information
- Other details:
 - Rectification (ReLU) non-linearity
 - 5 max-pool layers (x2 reduction)
 - no normalization
 - 3 fully-connected (FC) layers



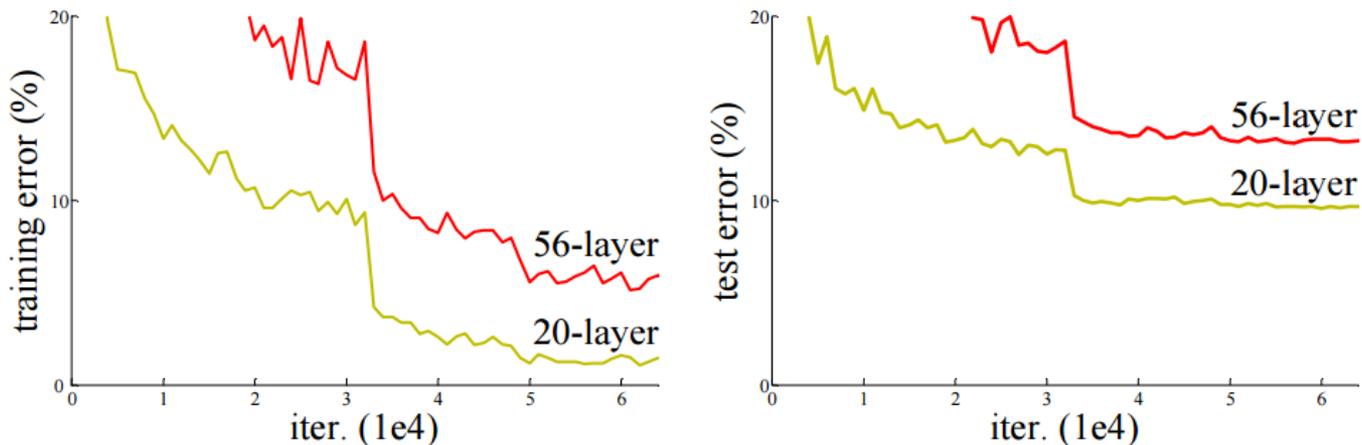
VGG-Net

- Why 3x3 layers?
 - Stacked conv. layers have a large receptive field
 - two 3x3 layers – 5x5 receptive field
 - three 3x3 layers – 7x7 receptive field
- More non-linearity
 - Less parameters to learn
 - ~140M per net

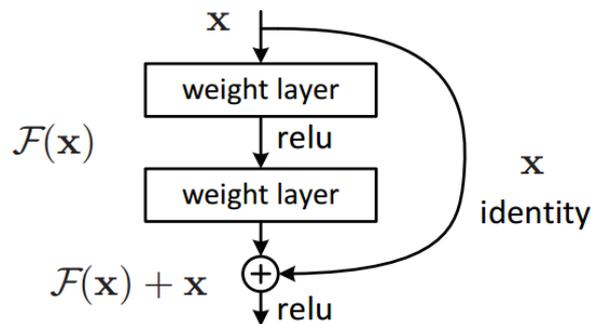


ResNet

- Can we just increase the #layer?



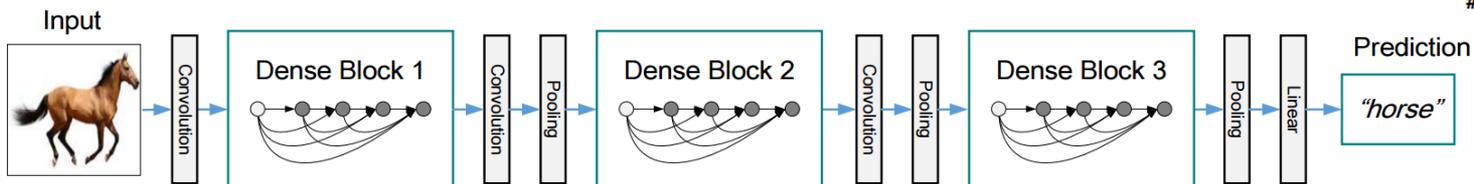
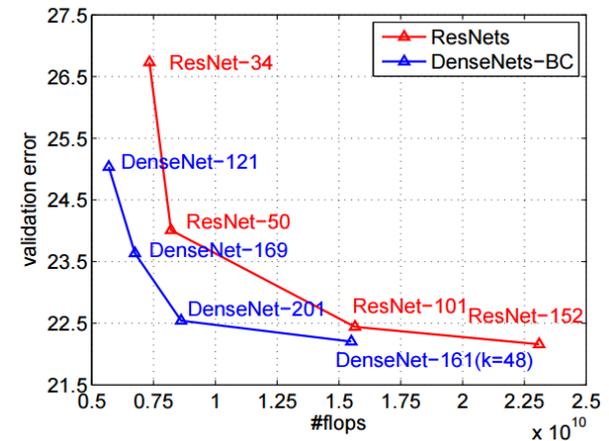
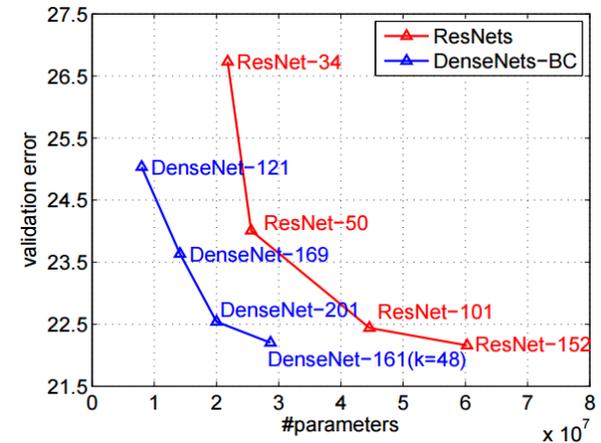
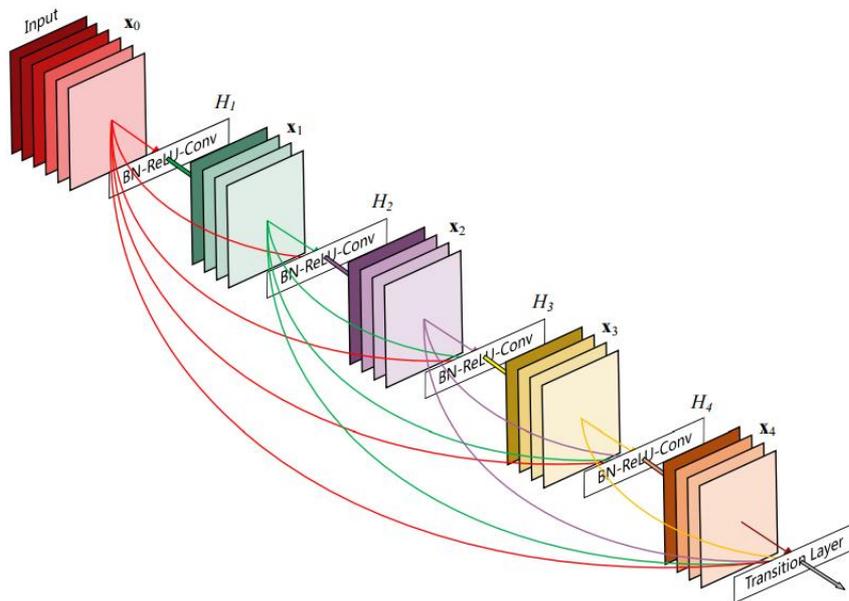
- How can we train very deep network?
 - Residual learning



method	top-5 err. (test)
VGG [41] (ILSVRC'14)	7.32
GoogLeNet [44] (ILSVRC'14)	6.66
VGG [41] (v5)	6.8
PReLU-net [13]	4.94
BN-inception [16]	4.82
ResNet (ILSVRC'15)	3.57

DenseNet

- Shorter connections (like ResNet) help
- Why not just connect them all?



Training Convolutional Neural Networks

- Backpropagation + stochastic gradient descent with momentum
 - [Neural Networks: Tricks of the Trade](#)
- Dropout
- Data augmentation
- Batch normalization
- Initialization
 - Transfer learning

Training CNN with gradient descent

- A CNN as composition of functions

$$f_{\mathbf{w}}(\mathbf{x}) = f_L(\dots (f_2(f_1(\mathbf{x}; \mathbf{w}_1); \mathbf{w}_2) \dots); \mathbf{w}_L)$$

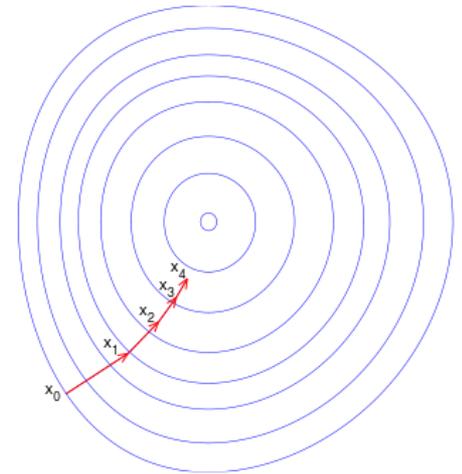
- Parameters

$$\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_L)$$

- Empirical loss function

$$L(\mathbf{w}) = \frac{1}{n} \sum_i l(z_i, f_{\mathbf{w}}(\mathbf{x}_i))$$

- Gradient descent



New weight

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta_t \frac{\partial f}{\partial \mathbf{w}}(\mathbf{w}^t)$$

Old weight

Learning rate

Gradient

An Illustrative example

$$f(x, y) = xy, \quad \frac{\partial f}{\partial x} = y, \quad \frac{\partial f}{\partial y} = x$$

Example: $x = 4, y = -3 \Rightarrow f(x, y) = -12$

Partial derivatives

$$\frac{\partial f}{\partial x} = -3, \quad \frac{\partial f}{\partial y} = 4$$

Gradient

$$\nabla f = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right]$$

$$f(x, y, z) = (x + y)z = qz$$

$$q = x + y$$

$$\frac{\partial q}{\partial x} = 1,$$

$$\frac{\partial q}{\partial y} = 1$$

$$f = qz$$

$$\frac{\partial f}{\partial q} = z,$$

$$\frac{\partial f}{\partial z} = q$$

Goal: compute the gradient

$$\nabla f = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right]$$

$$f(x, y, z) = (x + y)z = qz$$

$$q = x + y$$

$$\frac{\partial q}{\partial x} = 1,$$

$$\frac{\partial q}{\partial y} = 1$$

$$f = qz$$

$$\frac{\partial f}{\partial q} = z,$$

$$\frac{\partial f}{\partial z} = q$$

Chain rule:

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q} \frac{\partial q}{\partial x}$$

set some inputs

x = -2; y = 5; z = -4

perform the forward pass

q = x + y # q becomes 3

*f = q * z # f becomes -12*

perform the backward pass (backpropagation) in reverse order:

*# first backprop through f = q * z*

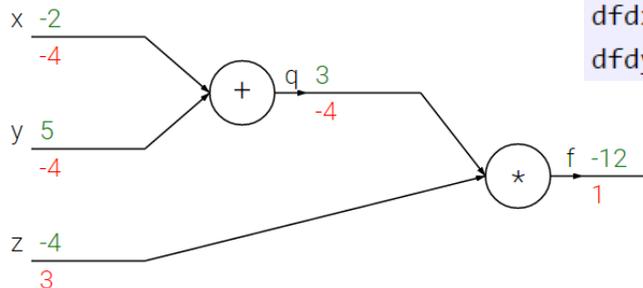
dfd_z = q # df/dz = q, so gradient on z becomes 3

dfd_q = z # df/dq = z, so gradient on q becomes -4

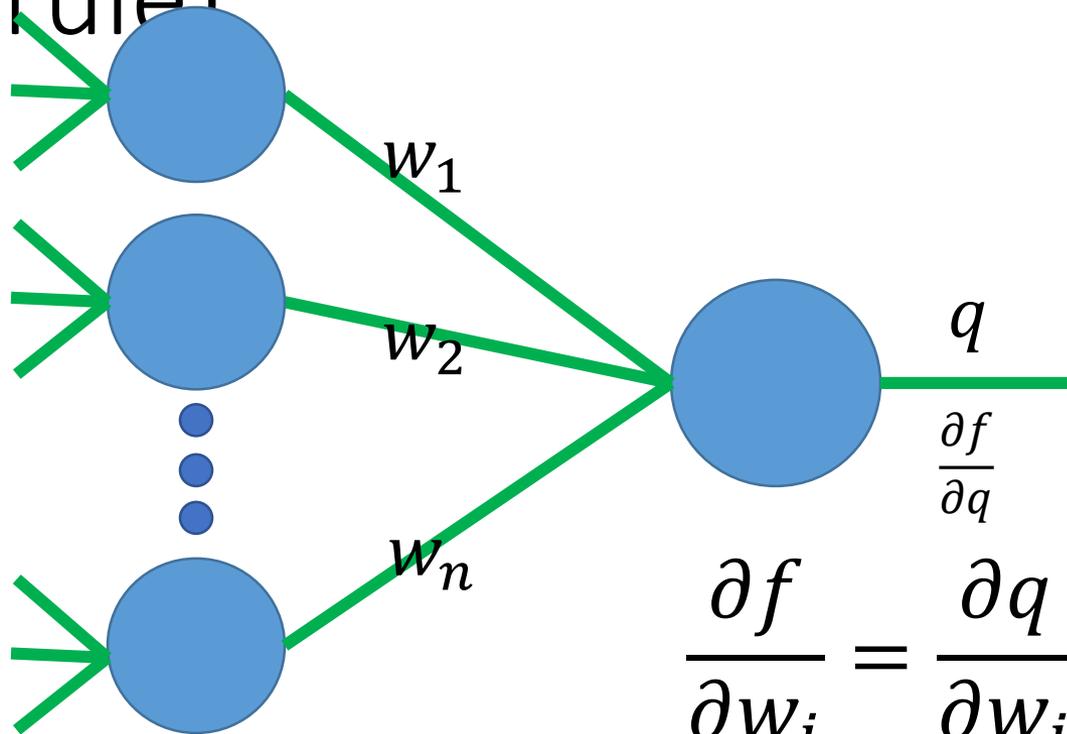
now backprop through q = x + y

*dfd_x = 1.0 * dfdq # dq/dx = 1. And the multiplication here is the chain rule!*

*dfd_y = 1.0 * dfdq # dq/dy = 1*



Backpropagation (recursive chain rule)



$$\frac{\partial f}{\partial w_i} = \frac{\partial q}{\partial w_i} \frac{\partial f}{\partial q}$$

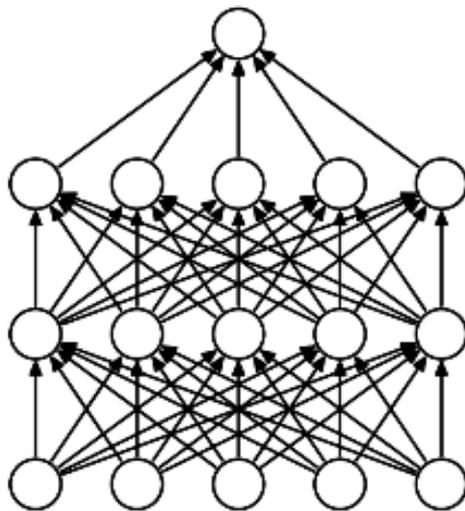
Local gradient

Gate gradient

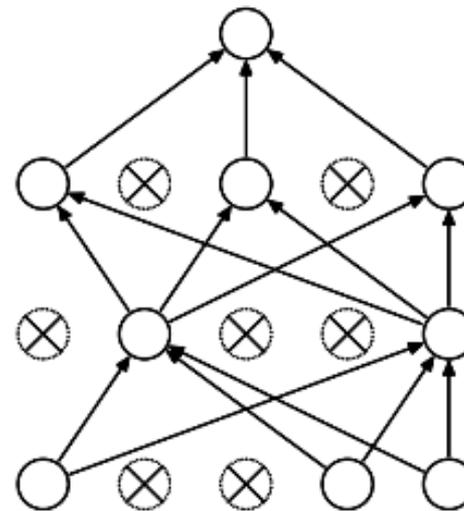
Can be computed during forward pass

The gate receives this during backprop

Dropout



(a) Standard Neural Net

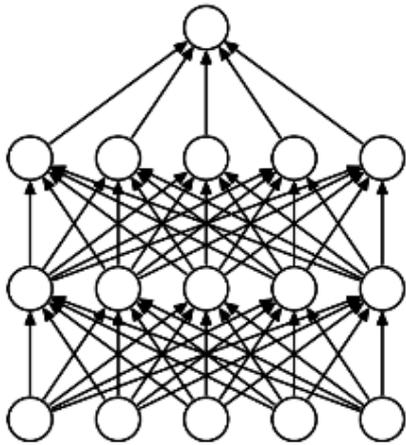


(b) After applying dropout.

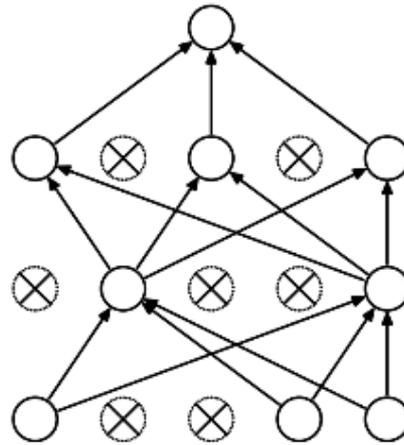
Intuition: successful conspiracies

- 50 people planning a conspiracy
- Strategy A: plan a big conspiracy involving 50 people
 - Likely to fail. 50 people need to play their parts correctly.
- Strategy B: plan 10 conspiracies each involving 5 people
 - Likely to succeed!

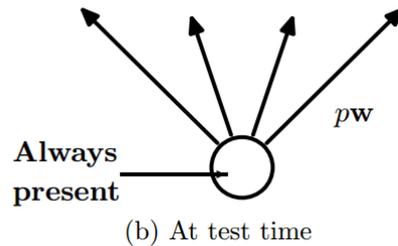
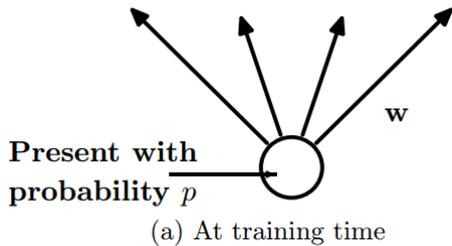
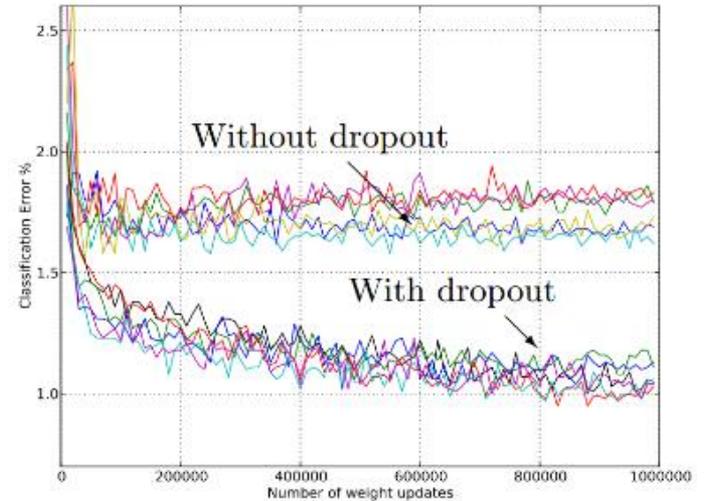
Dropout



(a) Standard Neural Net



(b) After applying dropout.



Main Idea: approximately combining exponentially many different neural network architectures efficiently

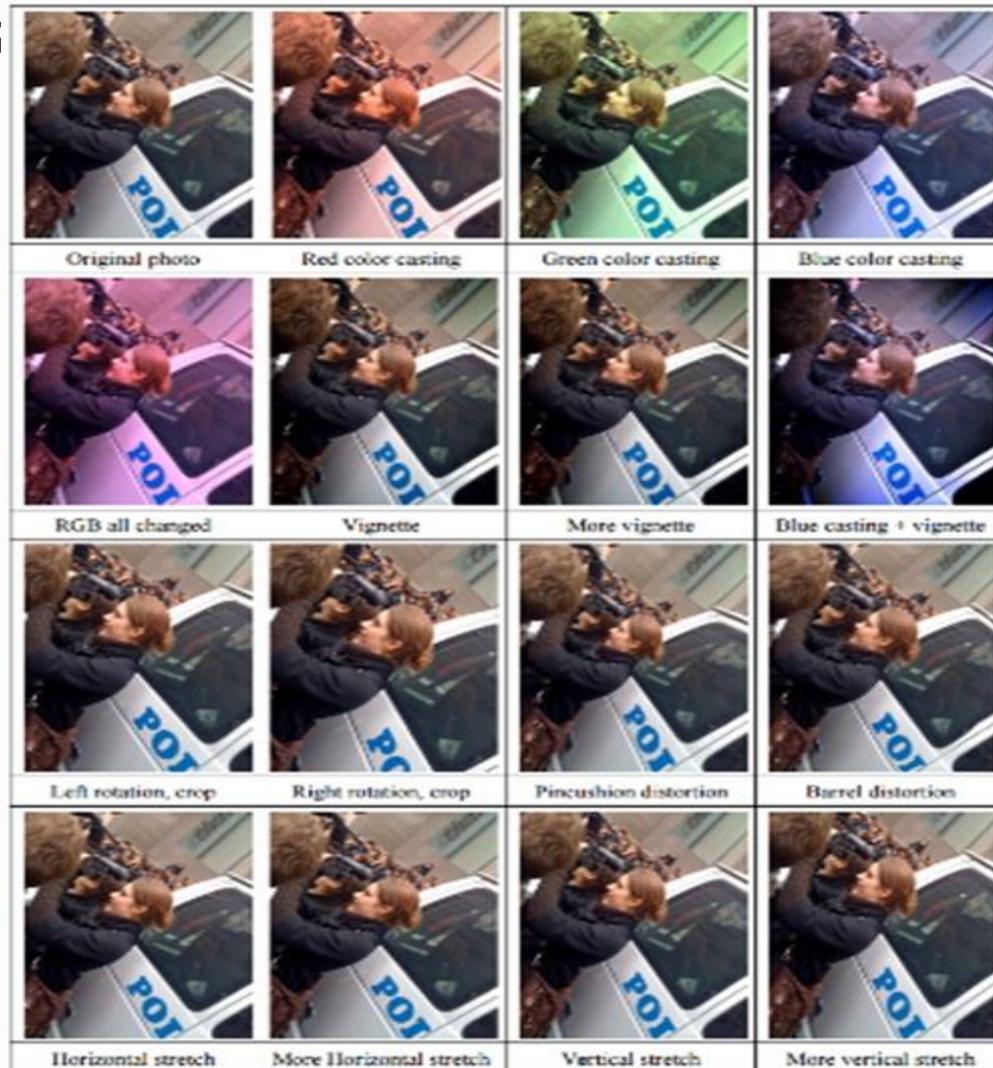
Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SVM on Fisher Vectors of Dense SIFT and Color Statistics	-	-	27.3
Avg of classifiers over FVs of SIFT, LBP, GIST and CSIFT	-	-	26.2
Conv Net + dropout (Krizhevsky et al., 2012)	40.7	18.2	-
Avg of 5 Conv Nets + dropout (Krizhevsky et al., 2012)	38.1	16.4	16.4

Table 6: Results on the ILSVRC-2012 validation/test set.

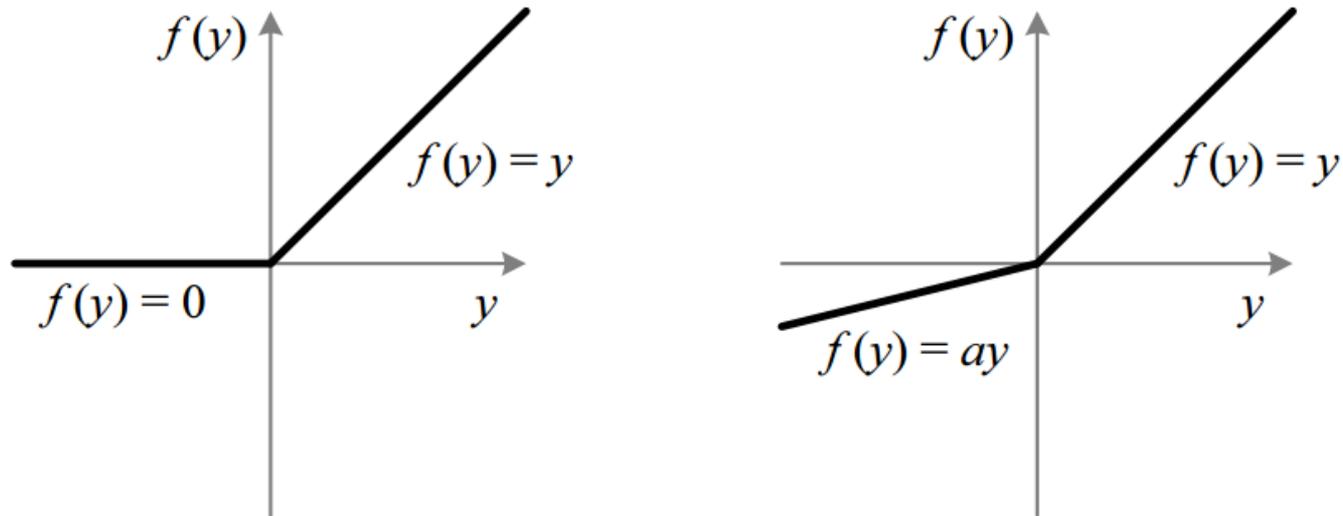
Dropout: A simple way to prevent neural networks from overfitting [[Srivastava JMLR 2014](#)]

Data Augmentation (Jittering)

- Create *virtual* training set
 - Horizontal flip
 - Random crop
 - Color casting
 - Geometric distortion



Parametric Rectified Linear Unit



	team	top-5 (test)
in competition ILSVRC 14	MSRA, SPP-nets [11]	8.06
	VGG [25]	7.32
	GoogLeNet [29]	6.66
post-competition	VGG [25] (arXiv v5)	6.8
	Baidu [32]	5.98
	MSRA, PReLU-nets	4.94

Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification [[He et al. 2015](#)]

Batch Normalization

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_1 \dots x_m\}$;

Parameters to be learned: γ, β

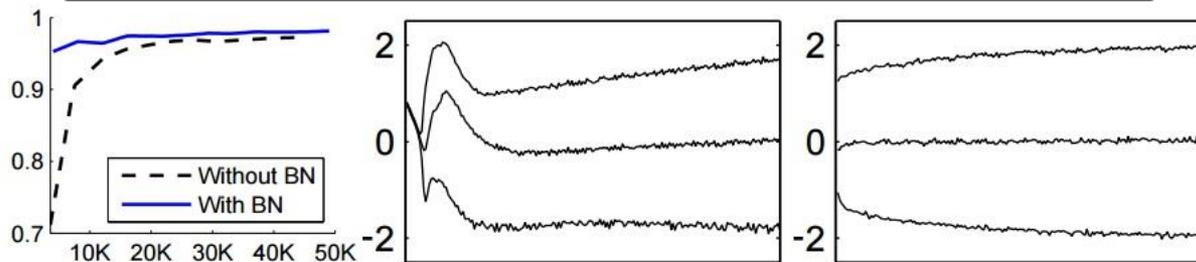
Output: $\{y_i = \text{BN}_{\gamma, \beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m} \sum_{i=1}^m x_i \quad // \text{ mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{\mathcal{B}})^2 \quad // \text{ mini-batch variance}$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \quad // \text{ normalize}$$

$$y_i \leftarrow \gamma \hat{x}_i + \beta \equiv \text{BN}_{\gamma, \beta}(x_i) \quad // \text{ scale and shift}$$



(a)

(b) Without BN

(c) With BN

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [[Ioffe and Szegedy 2015](#)]

Things to remember

- Visual categorization help transfer knowledge
- Convolutional neural networks
 - A cascade of conv + ReLU + pool
 - Representation learning
 - Advanced architectures
 - Tricks for training CNN