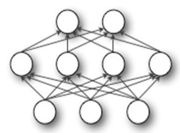
Deep learning for visual recognition

Thurs April 27
Kristen Grauman
UT Austin

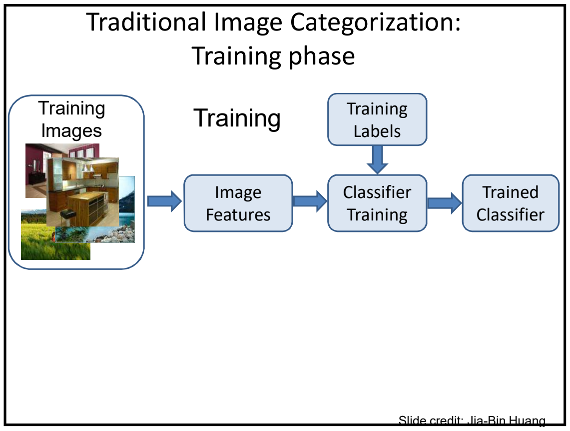


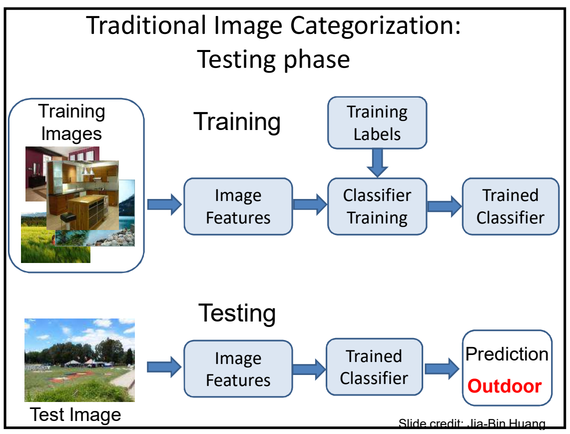
Last time

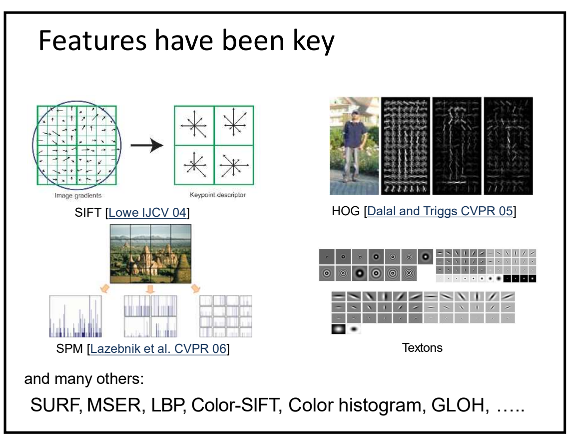
- Support vector machines (wrap-up)
  - Pyramid match kernels
- Evaluation
  - Scoring an object detector
  - Scoring a multi-class recognition system

Today

- (Deep) Neural networks
- Convolutional neural networks

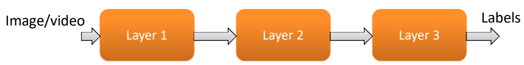

Traditional Image Categorization:
Training phase
Training Images
Training
Training Labels
Image Features
Classifier Training
Trained Classifier
Slide credit: Jia-Bin Huang


Traditional Image Categorization:
Testing phase
Training Images
Training
Training Labels
Image Features
Classifier Training
Trained Classifier
Testing
Image Features
Trained Classifier
Prediction
Outdoor
Test Image
Slide credit: Jia-Bin Huang


Features have been key
Image gradients
Keypoint descriptor
SIFT [Lowe IJCV 04]
HOG [Dalal and Triggs CVPR 05]
SPM [Lazebnik et al. CVPR 06]
Textons
and many others:
SURF, MSER, LBP, Color-SIFT, Color histogram, GLOH, …..

Learning a Hierarchy of Feature Extractors

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

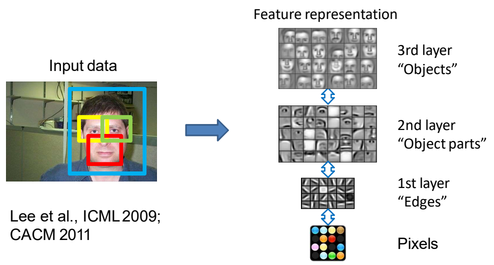


- Train all layers jointly

Slide: Rob Fergus

Learning Feature Hierarchy

Goal: Learn **useful higher-level features** from images



Lee et al., ICML 2009; CACM 2011

Slide: Rob Fergus

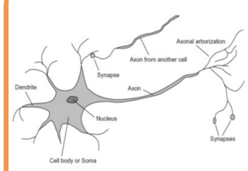
Learning Feature Hierarchy

- Better performance
- Other domains (unclear how to hand engineer):
  - Kinect
  - Video
  - Multi spectral

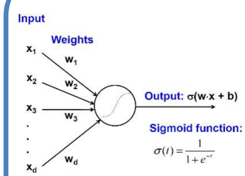
- Feature computation time
  - Dozens of features now regularly used [e.g., MKL]
  - Getting prohibitive for large datasets (10’s sec /image)

Slide: R. Fergus

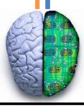
Biological neuron and Perceptrons



A biological neuron



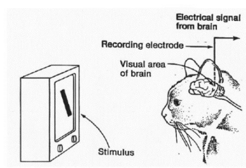
An artificial neuron (Perceptron)
- a linear classifier



Slide credit: Jia-Bin Huang

Simple, Complex and Hypercomplex cells

David H. Hubel and Torsten Wiesel

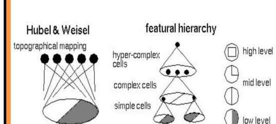


Suggested a **hierarchy** of **feature detectors** in the visual cortex, with higher level features responding to patterns of activation in lower level cells, and propagating activation upwards to still higher level cells.

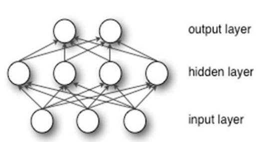
David Hubel's Eye, Brain, and Vision

Slide credit: Jia-Bin Huang

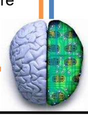
Hubel/Wiesel Architecture and Multi-layer Neural Network



Hubel and Weisel's architecture



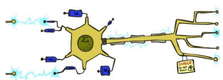
Multi-layer Neural Network
- A *non-linear* classifier



Slide credit: Jia-Bin Huang

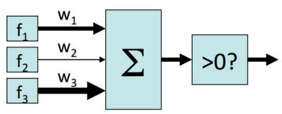
Neuron: Linear Perceptron

- Inputs are feature values
- Each feature has a weight
- Sum is the activation



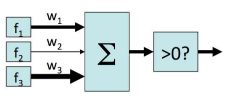
$$\text{activation}_w(x) = \sum_i w_i \cdot f_i(x) = w \cdot f(x)$$

- If the activation is:
  - Positive, output +1
  - Negative, output -1



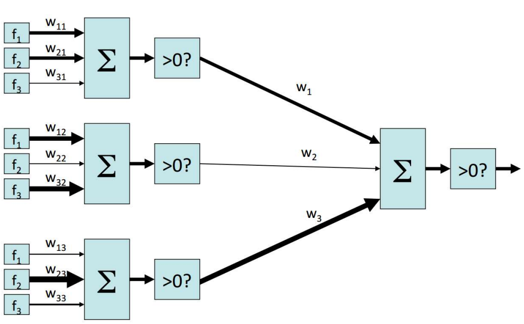
Slide credit: Pieter Abeel and Dan Klein

Two-layer perceptron network



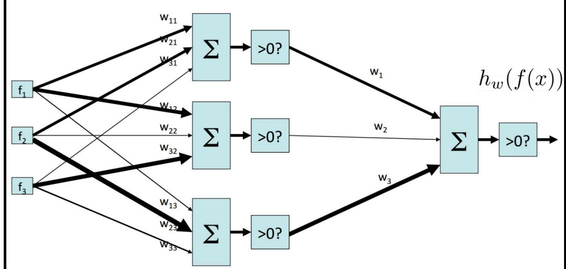
Slide credit: Pieter Abeel and Dan Klein

Two-layer perceptron network



Slide credit: Pieter Abeel and Dan Klein

Two-layer perceptron network



Slide credit: Pieter Abeel and Dan Klein

Learning w

- Training examples

$$(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})$$

- Objective: a misclassification loss

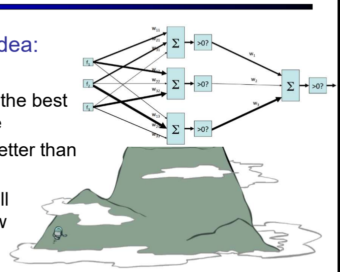
$$\min_{w} \sum_{i=1}^{m} \left(y^{(i)} - h_w(f(x^{(i)}))\right)^2$$

- Procedure:
  - Gradient descent / hill climbing

Slide credit: Pieter Abeel and Dan Klein

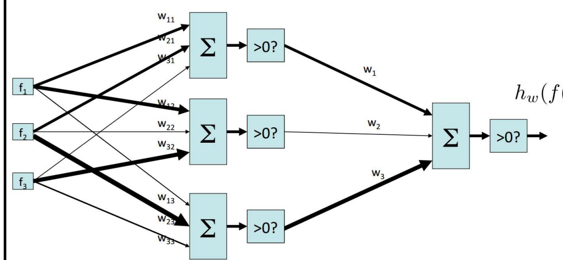
Hill climbing

- Simple, general idea:
  - Start wherever
  - Repeat: move to the best neighboring state
  - If no neighbors better than current, quit
  - Neighbors = small perturbations of w
- What's bad?
  - Complete?
  - Optimal?



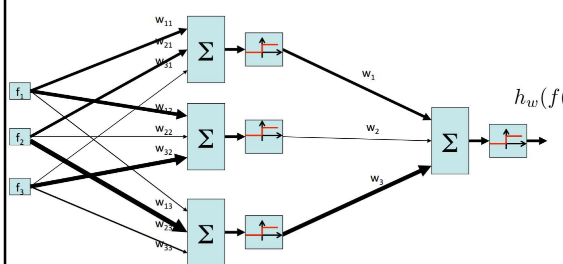
Slide credit: Pieter Abeel and Dan Klein

Two-layer perceptron network



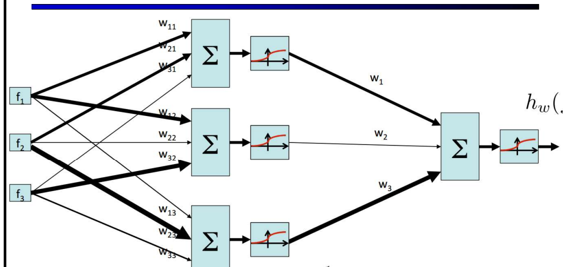
Slide credit: Pieter Abeel and Dan Klein

Two-layer perceptron network



Slide credit: Pieter Abeel and Dan Klein

Two-layer neural network



$$z \rightarrow \frac{1}{1+e^{-z}}$$

Slide credit: Pieter Abeel and Dan Klein

Neural network properties

- Theorem (Universal function approximators): A two-layer network with a sufficient number of neurons can approximate any continuous function to any desired accuracy
- Practical considerations:
  - Can be seen as learning the features
  - Large number of neurons
    - Danger for overfitting
  - Hill-climbing procedure can get stuck in bad local optima

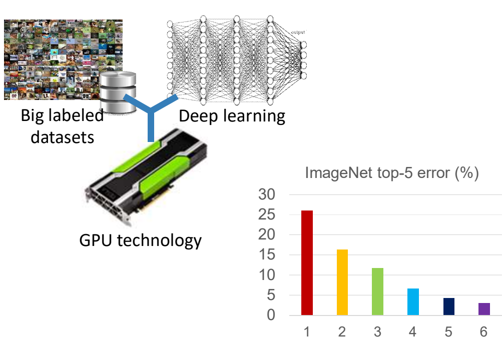
Approximation by Superpositions of Sigmoidal Function,1989

Slide credit: Pieter Abeel and Dan Klein

Today

- (Deep) Neural networks
- **Convolutional neural networks**

Significant recent impact on the field



Slide credit: Dinesh Jayaraman

Convolutional Neural Networks
(CNN, ConvNet, DCN)

- CNN = a multi-layer neural network with
  - **Local** connectivity:
    - Neurons in a layer are only connected to a small region of the layer before it
  - **Share** weight parameters across spatial positions:
    - Learning shift-invariant filter kernels

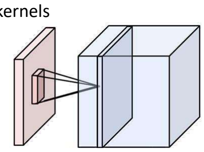


Image credit: A. Karpathy

Jia-Bin Huang and Derek Hoiem, UIUC

Neocognitron [Fukushima, Biological Cybernetics 1980]

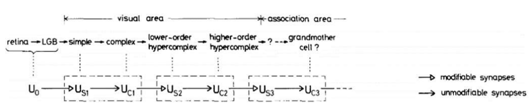
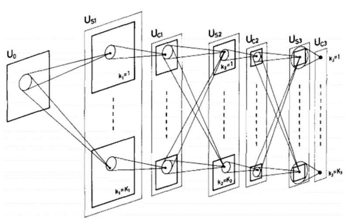


Fig. 1. Correspondence between the hierarchy model by Hubel and Wiesel, and the neural network of the neocognitron



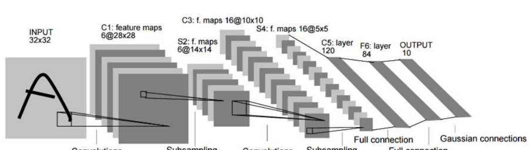
Deformation-Resistant Recognition

S-cells: (simple)
- extract local features

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

Jia-Bin Huang and Derek Hoiem, UIUC

LeNet [LeCun et al. 1998]



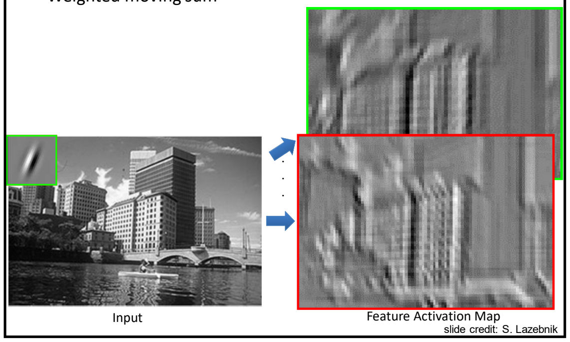
LeNet-1 from 1993

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

Jia-Bin Huang and Derek Hoiem, UIUC

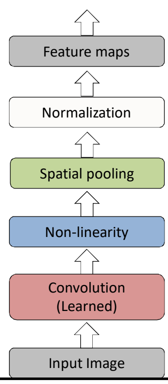
What is a Convolution?

- Weighted moving sum



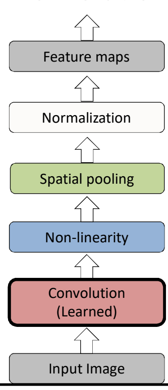
slide credit: S. Lazebnik

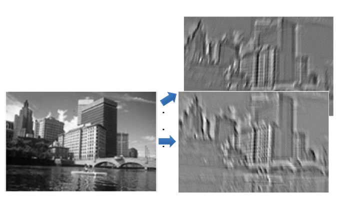
Convolutional Neural Networks



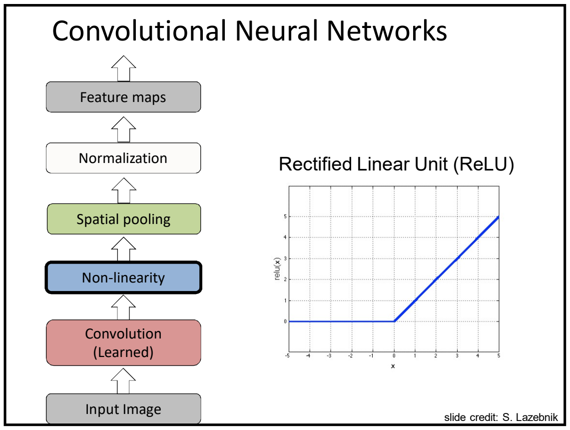
slide credit: S. Lazebnik

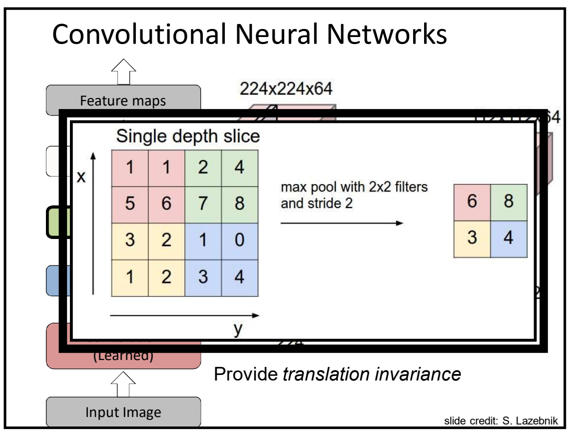
Convolutional Neural Networks

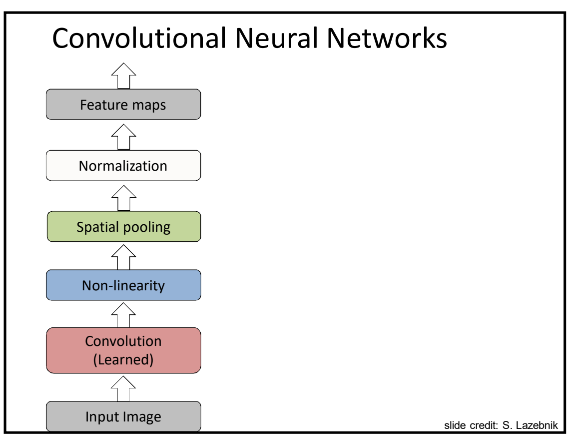




slide credit: S. Lazebnik

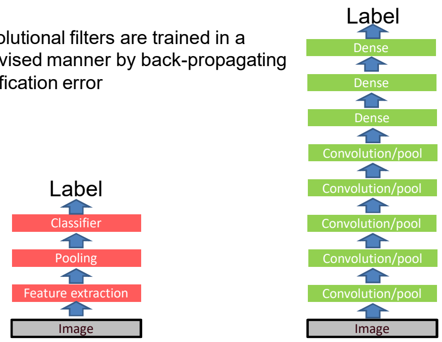

Convolutional Neural Networks
Feature maps
Normalization
Spatial pooling
Non-linearity
Convolution (Learned)
Input Image
Rectified Linear Unit (ReLU)
slide credit: S. Lazebnik


Convolutional Neural Networks
Feature maps
224x224x64
Single depth slice
x
y
1 1 2 4
5 6 7 8
3 2 1 0
1 2 3 4
max pool with 2x2 filters and stride 2
6 8
3 4
(Learned)
Provide translation invariance
Input Image
slide credit: S. Lazebnik


Convolutional Neural Networks
Feature maps
Normalization
Spatial pooling
Non-linearity
Convolution (Learned)
Input Image
slide credit: S. Lazebnik

Engineered vs. learned features

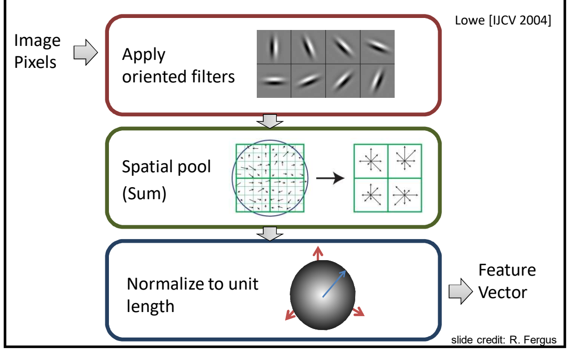
Convolutional filters are trained in a supervised manner by back-propagating classification error



Jia-Bin Huang and Derek Hoiem, UIUC

SIFT Descriptor

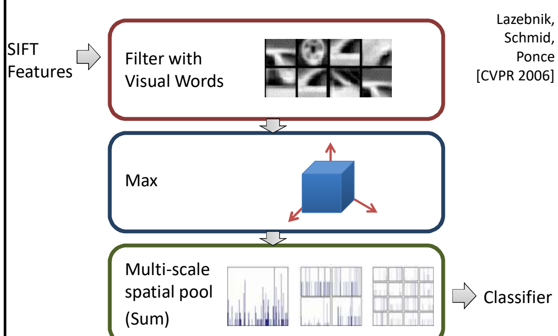
Lowe [IJCV 2004]



slide credit: R. Fergus

Spatial Pyramid Matching

Lazebnik, Schmid, Ponce [CVPR 2006]



slide credit: R. Fergus

Applications

- Handwritten text/digits
  - MNIST (0.17% error [Ciresan et al. 2011])
  - Arabic & Chinese [Ciresan et al. 2012]

- Simpler recognition benchmarks
  - CIFAR-10 (9.3% error [Wan et al. 2013])
  - Traffic sign recognition
    - 0.56% error vs 1.16% for humans [Ciresan et al. 2011]

Slide: R. Fergus

Application: ImageNet

[Deng et al. CVPR 2009]

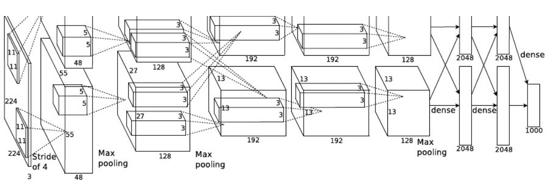
- ~14 million labeled images, 20k classes
- Images gathered from Internet
- Human labels via Amazon Turk

https://sites.google.com/site/deeplearningcvpr2014

Slide: R. Fergus

AlexNet

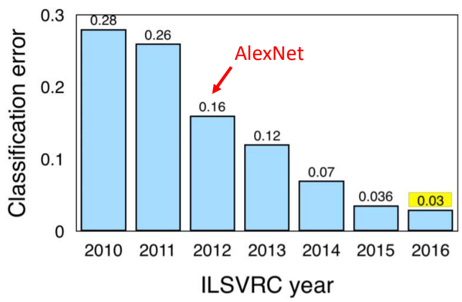
- Similar framework to LeCun'98 but:
  - Bigger model (7 hidden layers, 650,000 units, 60,000,000 params)
  - More data (10^6 vs. 10^3 images)
  - GPU implementation (50x speedup over CPU)
    - Trained on two GPUs for a week



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

Jia-Bin Huang and Derek Hoiem, UIUC

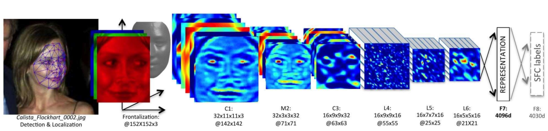
ImageNet Classification Challenge



http://image-net.org/challenges/talks/2016/ILSVRC2016_10_09_clsloc.pdf

Industry Deployment

- Used in Facebook, Google, Microsoft
- Image Recognition, Speech Recognition, ….
- Fast at test time



Taigman et al. DeepFace: Closing the Gap to Human-Level Performance in Face Verification, CVPR'14

Slide: R. Fergus

Beyond classification

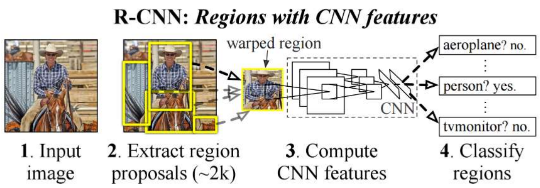
- Detection
- Segmentation
- Regression
- Pose estimation
- Matching patches
- Synthesis

and many more…

Jia-Bin Huang and Derek Hoiem, UIUC

R-CNN: Regions with CNN features

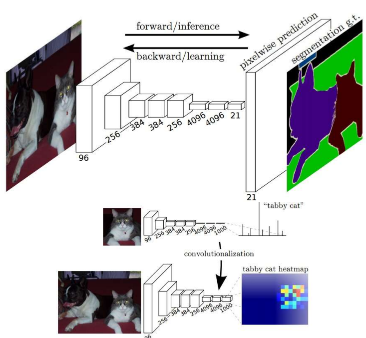
- Trained on ImageNet classification
- Finetune CNN on PASCAL



RCNN [Girshick et al. CVPR 2014]

Jia-Bin Huang and Derek Hoiem, UIUC

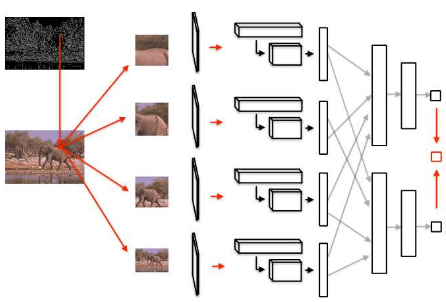
Labeling Pixels: Semantic Labels



Fully Convolutional Networks for Semantic Segmentation [Long et al. CVPR 2015]

Jia-Bin Huang and Derek Hoiem, UIUC

Labeling Pixels: Edge Detection



DeepEdge: A Multi-Scale Bifurcated Deep Network for Top-Down Contour Detection [Bertasius et al. CVPR 2015]

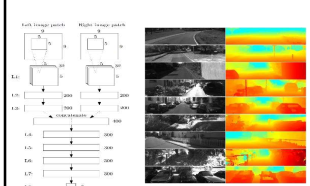
Jia-Bin Huang and Derek Hoiem, UIUC

CNN for Regression

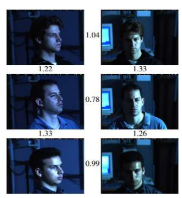
DeepPose [Toshev and Szegedy CVPR 2014]

Jia-Bin Huang and Derek Hoiem, UIUC

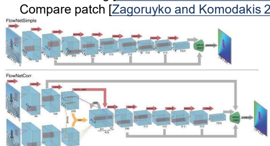
CNN as a Similarity Measure for Matching



Stereo matching [Zbontar and LeCun CVPR 2015]
Compare patch [Zagoruyko and Komodakis 2015]



FaceNet [Schroff et al. 2015]



FlowNet [Fischer et al 2015]

Match ground and aerial images
[Lin et al. CVPR 2015]

Jia-Bin Huang and Derek Hoiem, UIUC

Recap

- Neural networks / multi-layer perceptrons
  - View of neural networks as learning hierarchy of features
- Convolutional neural networks
  - Architecture of network accounts for image structure
  - "End-to-end" recognition from pixels
  - Together with big (labeled) data and lots of computation → major success on benchmarks, image classification and beyond