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 Labels
- Image Features
- Classifier Training
- Trained Classifier

Features have been key

- SIFT [Lowe IJCV 04]
- HOG [Dalal and Triggs CVPR 05]
- SPM [Lazebnik et al. CVPR 06]
- Textons
- SURF, MSER, LBP, Color-SIFT, Color histogram, GLOH, ….

Slide credit: Jia-Bin Huang

Traditional Image Categorization:
Testing phase

- Training Images
- Training Labels
- Image Features
- Classifier Training
- Trained Classifier
- Prediction: Outdoor

Slide credit: Jia-Bin Huang

Features have been key

- SIFT [Lowe IJCV 04]
- HOG [Dalal and Triggs CVPR 05]
- SPM [Lazebnik et al. CVPR 06]
- Textons
- SURF, MSER, LBP, Color-SIFT, Color histogram, GLOH, ….

Slide credit: Jia-Bin Huang
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

Learning Feature Hierarchy

Goal: Learn useful higher-level features from images

Input data

Feature representation

1st layer
“Edges”

2nd layer
“Object parts”

3rd layer
“Objects”

Lee et al., ICML 2009; CACM 2011

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)
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

Hubel/Wiesel Architecture and Multi-layer Neural Network

Hubel and Wiesel’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

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

Two-layer perceptron network
Learning $w$

- **Training examples**
  \[(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\]

- **Objective:** a misclassification loss
  \[
  \min_w \sum_{i=1}^{m} (y^{(i)} - h_w(f(x^{(i)})))^2
  \]

- **Procedure:**
  - Gradient descent / hill climbing

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?
Two-layer perceptron network

Two-layer neural network

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

Slide credit: Pieter Abeel and Dan Klein

Approximation by Superpositions of Sigmoidal Function, 1989

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

**Neocognitron** [Fukushima, Biological Cybernetics 1980]

- Deformation-Resistant Recognition
  - S-cells: (simple)
    - extract local features
  - C-cells: (complex)
    - allow for positional errors

Image credit: Jia-Bin Huang and Derek Hoiem, UIUC

**LeNet** [LeCun et al. 1998]

- Gradient-based learning applied to document recognition [LeCun, Bottou, Bengio, Haffner 1998]
- LeNet-1 from 1993

Image credit: Jia-Bin Huang and Derek Hoiem, UIUC
What is a Convolution?

• Weighted moving sum

Convolutional Neural Networks

Feature maps
Normalization
Spatial pooling
Non-linearity
Convolution (Learned)
Input Image

Feature maps
Normalization
Spatial pooling
Non-linearity
Convolution (Learned)
Input Image
Convolutional Neural Networks

- Feature maps
- Normalization
- Spatial pooling
- Non-linearity
- Convolution (Learned)
- Input Image

Rectified Linear Unit (ReLU)

Convolutional Neural Networks

Max-pooling: a non-linear down-sampling
Provide translation invariance
Engineered vs. learned features

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

SIFT Descriptor

Image Pixels

Apply oriented filters

Spatial pool (Sum)

Normalize to unit length

Feature Vector

SIFT Features

Filter with Visual Words

Max

Multi-scale spatial pool (Sum)

Classifier

Spatial Pyramid Matching

Lausen, Schmid, Ponce [CVPR 2006]

Jia-Bin Huang and Derek Hoiem, UIUC

Lowe [ICCV 2004]

slide credit: R. Fergus

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]

Application: ImageNet

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

AlexNet

- Similar framework to LeCun’98 but:
  - Bigger model (7 hidden layers, 650,000 units, 60,000,000 params)
  - More data (10^6 vs. 10^4 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

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

Slide: R. Fergus
ImageNet Classification Challenge

![ImageNet Classification Challenge Graph](http://image-net.org/challenges/talks/2016/ILSVRC2016_10_09_vldoc.pdf)

Industry Deployment

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

![DeepFace Example](http://image-net.org/challenges/talks/2016/ILSVRC2016_10_09_vldoc.pdf)

Tagman 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...
R-CNN: Regions with CNN features
• Trained on ImageNet classification
• Finetune CNN on PASCAL

Labeling Pixels: Semantic Labels

Labeling Pixels: Edge Detection
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