Deep learning for visual recognition

Tues April 23
Kristen Grauman
UT Austin

Last time

• Supervised classification continued
  • Nearest neighbors
  • Support vector machines
  • HoG pedestrians example
  • Kernels
  • Multi-class from binary classifiers

Recall: Examples of kernel functions

• Linear: \( K(x_i, x_j) = x_i^T x_j \)

• Gaussian RBF: \( K(x_i, x_j) = \exp\left(-\frac{|x_i - x_j|^2}{2\sigma^2}\right) \)

• Histogram intersection:
  \( K(x_i, x_j) = \sum \min(x_i(k), x_j(k)) \)

• Kernels go beyond vector space data
• Kernels also exist for "structured" input spaces like sets, graphs, trees…
Discriminative classification with sets of features?
- Each instance is an unordered set of vectors
- Varying number of vectors per instance

Partially matching sets of features

Optimal match: $O(m^3)$
Greedy match: $O(m^2 \log m)$
Pyramid match: $O(m)$

(Previous work: Indyk & Thaper, Bartal, Charikar, Agarwal & Varadarajan, …)

Pyramid match: main idea

Feature space partitions serve to “match” the local descriptors within successively wider regions.
Pyramid match: main idea

- Histogram intersection counts the number of possible matches at a given partitioning.

Pyramid match

\[ K_\Delta(X,Y) = \sum_{i=0}^{L} 2^{-i} \left( I(H_X^{(i)}, H_Y^{(0)}) - I(H_X^{(i)}, H_Y^{(i-1)}) \right) \]

- Measures the difficulty of a match at level \( i \).
- Number of newly matched pairs at level \( i \).

- For similarity, weights inversely proportional to bin size (or may be learned).
- Normalize these kernel values to avoid favoring large sets.

Pyramid match

- Optimal partial matching:
  - \( O(m^2) \)
  - \( O(mL) \)

BoW Issue:
No spatial layout preserved!

Too much?

Too little?

---

Spatial pyramid match
- Make a pyramid of bag-of-words histograms.
- Provides some loose (global) spatial layout information

\[
K^L(X, Y) = \sum_{m=1}^{M} \kappa^L(X_m, Y_m)
\]

Sum over PMKs computed in image coordinate space, one per word.

[BoW credit: Kristen Grauman]

[Lazebnik, Schmid & Ponce, CVPR 2006]

[BoW credit: Kristen Grauman]

Spatial pyramid match
- Make a pyramid of bag-of-words histograms.
- Provides some loose (global) spatial layout information

[Lazebnik, Schmid & Ponce, CVPR 2006]
Spatial pyramid match

- Can capture scene categories well—texture-like patterns but with some variability in the positions of all the local pieces.
- Sensitive to global shifts of the view

Today

- (Deep) Neural networks
- Convolutional neural networks
Traditional Image Categorization:
Training phase

Traditional Image Categorization:
Testing phase

Features have been key

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

---

**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"
- Pixels
- Labels

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 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 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

Two-layer perceptron network

Two-layer perceptron network

Two-layer perceptron network
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?
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 credits: Pieter Abeel and Dan Klein

Approximation by Superpositions of Sigmoidal Functions, 1989

Today

- (Deep) Neural networks
- Convolutional neural networks

Significant recent impact on the field

- Big labeled datasets
- Deep learning
- GPU technology

ImageNet top-5 error (%)
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

LeNet [LeCun et al. 1998]

What is a Convolution?

- Weighted moving sum
Convolutional Neural Networks

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

[Diagram of Convolutional Neural Networks]

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

[Diagram of Convolutional Neural Networks with Rectified Linear Unit (ReLU)]

- Rectified Linear Unit (ReLU)

[Diagram of Convolutional Neural Networks with Slide Credit: S. Lazebnik]
Convolutional Neural Networks

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

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

Max pooling: a non-linear down-sampling to provide translation invariance.

Engineered vs. learned features

Jia-Bin Huang and Derek Hoiem, UIUC
SIFT Descriptor

Image Pixels → Apply oriented filters → Spatial pool (Sum) → Normalize to unit length → Feature Vector

Lowe [IJCV 2004]

Spatial Pyramid Matching

SIFT Features → Filter with Visual Words → Multi-scale spatial pool (Sum) → Classifier

Lazebnik, Schmid, Ponce [CVPR 2006]

Visualizing what was learned

• What do the learned filters look like?

Typical first layer filters
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. 10^3 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


Industry Deployment

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

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