This Lecture: Advanced Machine Learning

- **Regression Approaches**
  - Linear Regression
  - Regularization (Ridge, Lasso)
  - Kernels (Kernel Ridge Regression)
  - Gaussian Processes

- **Approximate Inference**
  - Sampling Approaches
  - MCMC

- **Deep Learning**
  - Linear Discriminants
  - Neural Networks
  - Backpropagation & Optimization
  - CNNs, RNNs, ResNets, etc.

Topics of This Lecture

- Recap: CNNs
- CNN Architectures
  - LeNet
  - AlexNet
  - VGGNet
  - GoogLeNet
  - ResNets
- Visualizing CNNs
  - Visualizing CNN features
  - Visualizing responses
  - Visualizing learned structures
- Applications

Recap: Convolutional Neural Networks

- Neural network with specialized connectivity structure
  - Stack multiple stages of feature extractors
  - Higher stages compute more global, more invariant features
  - Classification layer at the end

Recap: Intuition of CNNs

- Convolutional net
  - Share the same parameters across different locations
  - Convolutions with learned kernels
- Learn multiple filters
  - E.g. 1000x1000 image
    - 100 filters
    - 10x10 filter size
  - Only 10k parameters
- Result: Response map
  - Size: 1000x1000x100
  - Only memory, not params!

Recap: Convolution Layers

- All Neural Net activations arranged in 3 dimensions
  - Multiple neurons all looking at the same input region, stacked in depth
  - Form a single [1x1xdepth] depth column in output volume.
Recap: Activation Maps

Each activation map is a depth slice through the output volume.

Recap: Pooling Layers

- Effect:
  - Make the representation smaller without losing too much information
  - Achieve robustness to translations

Topics of This Lecture

- Recap: CNNs
- CNN Architectures
  - LeNet
  - AlexNet
  - VGGNet
  - GoogLeNet
- Visualizing CNNs
  - Visualizing CNN features
  - Visualizing responses
  - Visualizing learned structures
- Applications

Recap: ImageNet Challenge 2012

- ImageNet
  - ~14M labeled internet images
  - 20k classes
  - Human labels via Amazon Mechanical Turk
- Challenge (ILSVRC)
  - 1.2 million training images
  - 1000 classes
  - Goal: Predict ground-truth class within top-5 responses
  - Currently one of the top benchmarks in Computer Vision

CNN Architectures: AlexNet (2012)

- Similar framework as LeNet, but
  - Bigger model (7 hidden layers, 650k units, 60M parameters)
  - More data (10^7 images instead of 10^3)
  - GPU implementation
  - Better regularization and up-to-date tricks for training (Dropout)

ILSVRC 2012 Results

- AlexNet almost halved the error rate
  - 16.4% error (top-5) vs. 26.2% for the next best approach
  - A revolution in Computer Vision
  - Acquired by Google in Jan ‘13, deployed in Google+ in May ‘13
CNN Architectures: VGGNet (2014/15)

**Main ideas**
- Deeper network
- Stacked convolutional layers with smaller filters (+ nonlinearity)
- Detailed evaluation of all components

**Results**
- Improved ILSVRC top-5 error rate to 6.7%.

---

Comparison: AlexNet vs. VGGNet

- **Receptive fields in the first layer**
  - AlexNet: $11 \times 11$, stride 4
  - Zeiler & Fergus: $7 \times 7$, stride 2
  - VGGNet: $3 \times 3$, stride 1

- **Why that?**
  - If you stack $3 \times 3$ on top of another $3 \times 3$ layer, you effectively get a $5 \times 5$ receptive field.
  - With three $3 \times 3$ layers, the receptive field is already $7 \times 7$.
  - But much fewer parameters: $3 \times 3 \times 27$ instead of $7 \times 7 = 49$.
  - In addition, non-linearities in-between $3 \times 3$ layers for additional discriminativity.

---

GoogLeNet Visualization

Inception module + copies

Auxiliary classification outputs for training the lower layers (deprecated)

---

GoogLeNet (2014)

**Main ideas**
- “Inception” module as modular component
- Learns filters at several scales within each module


---

Results on ILSVRC

- **VGGNet and GoogLeNet perform at similar level**
  - Comparison: human performance ~5% [Karpathy]
Newest Development: Residual Networks

- Core component
  - Skip connections bypassing each layer
  - Better propagation of gradients to the deeper layers
  - We’ll analyze this mechanism in more detail later...

Understanding the ILSVRC Challenge

- Imagine the scope of the problem!
  - 1000 categories
  - 1.2M training images
  - 50k validation images

- This means...
  - Speaking out the list of category names at 1 word/s...
    - ...takes 15mins.
  - Watching a slideshow of the validation images at 2s/image...
    - ...takes a full day (24h+).
  - Watching a slideshow of the training images at 2s/image...
    - ...takes a full month.
Quirks and Limitations of the Data Set

- Generated from WordNet ontology
  - Some animal categories are overrepresented
  - E.g., 120 subcategories of dog breeds
⇒ 6.7% top-5 error looks all the more impressive

Visualizing CNNs

- Recap: CNNs
- CNN Architectures
  - LeNet
  - AlexNet
  - VGGNet
  - GoogLeNet
- Visualizing CNNs
  - Visualizing CNN features
  - Visualizing responses
  - Visualizing learned structures
- Applications
What Does the Network React To?

- Occlusion Experiment
  - Mask part of the image with an
    occluding square.
  - Monitor the output

Input image

Total activation in most active 5th layer feature map
Other activations from the same feature map.

Input image

Total activation in most active 5th layer feature map
Other activations from the same feature map.
What Does the Network React To?

- Total activation in most active 5th layer feature map
- Other activations from the same feature map.

Inceptionism: Dreaming ConvNets

- Idea
  - Start with a random noise image.
  - Enhance the input image such as to enforce a particular response (e.g., banana).
  - Combine with prior constraint that image should have similar statistics as natural images.
  - Network hallucinates characteristics of the learned class.

Inceptionism: Dreaming ConvNets

- Results

http://googleresearch.blogspot.de/2015/06/inceptionism-going-deeper-into-neural.html

Inceptionism: Dreaming ConvNets

http://googleresearch.blogspot.de/2015/07/deepdream-code-example-for-visualizing.html

https://www.youtube.com/watch?v=IREsxuWQ0ig

Topics of This Lecture

- Recap: CNNs
- CNN Architectures
  - LeNet
  - AlexNet
  - VGGNet
  - GoogLeNet
- Visualizing CNNs
  - Visualizing CNN features
  - Visualizing responses
  - Visualizing learned structures
- Applications

The Learned Features are Generic

- Experiment: feature transfer
  - Train network on ImageNet
  - Chop off last layer and train classification layer on CalTech256
  - State of the art accuracy already with only 6 training images

state of the art level (pre-CNN)
Other Tasks: Detection

R-CNN: Regions with CNN features

1. Input image
2. Extract region proposals (~2k)
3. Compute CNN features
4. Classify regions

- Results on PASCAL VOC Detection benchmark
  - Pre-CNN state of the art: 35.1% mAP [Uijlings et al., 2013]
  - 33.4% mAP DPM
  - R-CNN: 53.7% mAP


Faster R-CNN (based on ResNets)


Faster R-CNN (based on ResNets)


Other Tasks: Semantic Segmentation

[B. Leibe]

Semantic Segmentation

[Pohlen, Hermans, Mathias, Leibe, arXiv 2016]

- More recent results
  - Based on an extension of ResNets

Other Tasks: Face Verification

Y. Taigman, M. Yang, M. Ranzato, L. Wolf, DeepFace: Closing the Gap to Human-Level Performance in Face Verification, CVPR 2014

Slide credit: Svetlana Lazebnik
Commercial Recognition Services

- E.g., clarifai

Try it out with your own media

Upload an image or video file under 10MB or give us a direct link to a file on the web.

Be careful when taking test images from Google Search.

- Chances are they may have been seen in the training set...

References and Further Reading

- LeNet

- AlexNet

- VGGNet

- GoogLeNet

- ResNet

Effect of Multiple Convolution Layers

Feature visualisation of convolutional net trained on ImageNet from [Zeller & Fergus 2013]