Lecture 34: Datasets

Visual Object Classes Challenge 2009 (VOC2009)
Data Sets

• Critical to the success of deep learning
  – Object classification and segmentation
  – Scene classification
  – Materials

• Examples
  – PASCAL VOC
    • Not Crowdsourced, bounding boxes, 20 categories
  – ImageNet
    • Huge, Crowdsourced, Hierarchical, Iconic objects
  – SUN Scene Database
    • Not Crowdsourced, 397 (or 720) scene categories
  – Microsoft COCO
    • Crowdsourced, large
  – Material Database: OpenSurfaces
The PASCAL Visual Object Classes Challenge 2009 (VOC2009)

- Twenty object categories (aeroplane to TV/monitor)

- Three challenges:
  - Classification challenge (is there an X in this image?)
  - Detection challenge (draw a box around every X)
  - Segmentation challenge (which class is each pixel?)
Dataset: Collection

- Images downloaded from **flickr**
  - 500,000 images downloaded and random subset selected for annotation
Examples

Aeroplane  Bicycle  Bird  Boat  Bottle

Bus  Car  Cat  Chair  Cow
Examples

Dining Table  Dog  Horse  Motorbike  Person

Potted Plant  Sheep  Sofa  Train  TV/Monitor
Classification Challenge

- Predict whether at least one object of a given class is present in an image

is there a cat?
Precision = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}

Recall = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}
Precision in red, recall in yellow

https://uberpython.wordpress.com/2012/01/01/precision-recall-sensitivity-and-specificity/
Precision Recall curves

- Related to but different from ROC curves
- Start at (0, 1), higher curves are better

- Average Precision (AP) = area under the curve
**Max AP: 88.1% (aeroplane) ... 40.8% (potted plant)**
Pascal VOC 2012 Average Precision
Ranked Images: Aeroplane

- Class images: Highest ranked
- Class images: Lowest ranked
- Non-class images: Highest ranked
- Context?
Detection Challenge

- Predict the bounding boxes of all objects of a given class in an image (if any)
Evaluating Bounding Boxes

- Area of Overlap (AO) Measure

\[ AO(B_{gt}, B_p) = \frac{|B_{gt} \cap B_p|}{|B_{gt} \cup B_p|} \]
AP by Class

Chance essentially 0
**PASCAL VOC 2005-2012**

- **20 object classes**
- **22,591 images**

**Classification:** person, motorcycle

**Detection**

**Segmentation**

**Action:** riding bicycle

Large Scale Visual Recognition Challenge (ILSVRC) 2010-2012

20 object classes | 22,591 images
1000 object classes | 1,431,167 images

Variety of object classes in ILSVRC

**PASCAL**
- birds: bird
- bottles: bottle
- cars: car

**ILSVRC**
- birds: flamingo, cock, ruffed grouse, quail, partridge
- bottles: pill bottle, beer bottle, wine bottle, water bottle, pop bottle
- cars: race car, wagon, minivan, jeep, cab
Variety of object classes in ILSVRC

- **Amount of Texture**
  - Screwdriver, Hatchet, Ladybug, Honeycomb

- **Color Distinctiveness**
  - Coffee mug, Cleaver, Bagel, Red Wine

- **Shape Distinctiveness**
  - Jigsaw Puzzle, Foreland, Lipstick, Bell

- **Real-world Size**
  - Orange, Mask, Parachute, Airliner
How do we classify scenes?

Different objects, different spatial layout
Which are the important elements?

Similar objects, and similar spatial layout

Different lighting, different materials, different “stuff”
Scene Categorization

Oliva and Torralba, 2001

Coast  Forest  Highway  Inside City  Mountain  Open Country  Street  Tall Building

Fei Fei and Perona, 2005

Bedroom  Kitchen  Living Room  Office  Suburb

Lazebnik, Schmid, and Ponce, 2006

Industrial  Store

15 Scene Database
SUN Database: Large-scale Scene Categorization and Detection

Jianxiong Xiao, James Hays†, Krista A. Ehinger, Aude Oliva, Antonio Torralba
Massachusetts Institute of Technology
† Brown University
397 Well-sampled Categories
<table>
<thead>
<tr>
<th>Scene category</th>
<th>Most confusing categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inn (0%)</td>
<td>Restaurant patio (44%)</td>
</tr>
<tr>
<td>Bayou (0%)</td>
<td>River (67%)</td>
</tr>
<tr>
<td>Basilica (0%)</td>
<td>Cathedral (29%)</td>
</tr>
<tr>
<td></td>
<td>Chalet (19%)</td>
</tr>
<tr>
<td></td>
<td>Coast (8%)</td>
</tr>
<tr>
<td></td>
<td>Courthouse (21%)</td>
</tr>
</tbody>
</table>
Now it’s the era of Big Data and Deep Learning

- **Places Database**
- ~7 million images from 476 scene categories
### ImageNet-CNN and Places-CNN

- Same structure as AlexNet, but trained on different databases.

<table>
<thead>
<tr>
<th></th>
<th>SUN397</th>
<th>MIT Indoor67</th>
<th>Scene15</th>
<th>SUN Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Places-CNN feature</td>
<td>54.32±0.14</td>
<td>68.24</td>
<td>90.19±0.34</td>
<td>91.29</td>
</tr>
<tr>
<td>ImageNet-CNN feature</td>
<td>42.61±0.16</td>
<td>56.79</td>
<td>84.23±0.37</td>
<td>89.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Caltech101</td>
<td>Caltech256</td>
<td>Action40</td>
<td>Event8</td>
</tr>
<tr>
<td>Places-CNN feature</td>
<td>65.18±0.88</td>
<td>45.59±0.31</td>
<td>42.86±0.25</td>
<td>94.12±0.99</td>
</tr>
<tr>
<td>ImageNet-CNN feature</td>
<td>87.22±0.92</td>
<td>67.23±0.27</td>
<td>54.92±0.33</td>
<td>94.42±0.76</td>
</tr>
</tbody>
</table>
Microsoft COCO

We present a new dataset with the goal of advancing the state-of-the-art in object recognition by placing the question of object recognition in the context of the broader question of scene understanding. This is achieved by gathering images of complex everyday scenes containing common objects in their natural context. Objects are labeled using per-instance segmentations to aid in precise object localization. Our dataset contains photos of 91 objects types that would be easily recognizable by a 4 year old. With a total of 2.5 million labeled instances in 328k images, the creation of our dataset drew upon extensive crowd worker involvement via novel user interfaces for category detection.

instance spotting
correlation to PASCAL
bounding box and

(a) Image classification

(b) Object localization

(c) Semantic segmentation

(d) This work

✓ Instance segmentation
✓ Non-iconic Images
Fig. 2: Example of (a) iconic object images, (b) iconic scene images, and (c) non-iconic images.
Fig. 3: Our annotation pipeline is split into 3 primary tasks: (a) labeling the categories present in the image (§4.1), (b) locating and marking all instances of the labeled categories (§4.2), and (c) segmenting each object instance (§4.3).
Material Database

• Different domain
  – Most of the focus has been on objects
  – Our focus on materials

• Sean Bell, Paul Upchurch, Noah Snavely, Kavita Bala
  – OpenSurfaces [2013]
  – Segmentation interface used by Microsoft COCO
OpenSurfaces

Get novice workers to accurately describe material appearance in [scalable, verifiable, and economical] way

context: “kitchen”
object: “countertop”

Material: “granite”, diffuse, specular, roughness

Texture

- Open Surfaces: opensurfaces.cs.cornell.edu
Pipeline Preview

1. Material segmentation: Draw boundaries
2. Name material
3. Reflectance
4. Texture
110,000 Segmentations
25,000 Textures
50,000 Reflectances
Data

• More is more....
Classification

(assume given set of discrete labels)
{dog, cat, truck, plane, ...}

→

cat
Localization

Model must output:

- **class** (integer)
- $x_1,y_1,x_2,y_2$ **bounding box** coordinates
**Very Deep Convolutional Networks for Large-Scale Image Recognition**, Simonyan et al., 2014

**OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks**, Sermanet et al., 2014

**Idea: train a Localization net**
Take out Softmax loss, swap in L2 (regression) loss, **fine-tune** the classification network.

swap the Softmax layer at the end with L2 loss
**Idea: train a Localization net**

Take out Softmax loss, swap in L2 (regression) loss, **fine-tune** the classification network.

predictions: instead of class scores, now interpreted as the 4 bounding box coords (also 4D vector from net)

$$L_i = \|f - y_i\|_2^2$$

targets: true bounding box 4D vector of $[x_1, y_1, x_2, y_2]$

swap the Softmax layer at the end with L2 loss
Very Deep Convolutional Networks for Large-Scale Image Recognition, Simonyan et al., 2014
OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks, Sermanet et al., 2014

In practice:

- It works better to predict a **4D vector for every class** (e.g. 4000D vector for 1000 ImageNet classes). During training only backprop the loss for the correct class.
- apply at **multiple locations and scales**

![Diagram of convolutional network architecture]

*swap the Softmax layer at the end with L2 loss*
• Heavily-modified Caffe C++ toolbox
• Multiple GPU support
  • 4 x NVIDIA Titan, off-the-shelf workstation
  • data parallelism for training and testing
  • ~3.75 times speed-up, 2-3 weeks for training
Summary of VGG

• Localisation task
  – 1\textsuperscript{st} place, 25.3% error

• Classification task
  – 2\textsuperscript{nd} place, 7.3% error

• Deep: 19 weight layers
Top-5 Localisation Error (Test Set)

- VGG: 25.3
- GoogLeNet: 26.4
- OverFeat (2013): 29.9
- SYSU: 31.9

Better
Detection

Needs to find all instances of the various classes

Model must output:

A set of detections

Each detection has:
- confidence
- class (integer)
- $x_1,y_1,x_2,y_2$ bounding box coordinates
Rich feature hierarchies for accurate object detection and semantic segmentation
[Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik]

Idea: Turn a Detection Problem into an Image Classification problem (but over image regions).

Content of every labeled bounding box for is a positive example for a class.

Every other bounding box in the image is a special negative class.
Rich feature hierarchies for accurate object detection and semantic segmentation
[Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik]

Idea: Turn a Detection Problem into an Image Classification problem (but over image regions).
Rich feature hierarchies for accurate object detection and semantic segmentation

[Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik]
Video Classification

Two-Stream Convolutional Networks for Action Recognition in Videos [Simonyan et al.], 2014

Long-term Recurrent Convolutional Networks for Visual Recognition and Description [Donahue et al.], 2014

Large-scale Video Classification with Convolutional Neural Networks [Karpathy et al.], 2014
CNN Features off-the-shelf: an Astounding Baseline for Recognition

Ali Sharif Razavian  Hossein Azizpour  Josephine Sullivan  Stefan Carlsson
CVAP, KTH (Royal Institute of Technology)
Stockholm, Sweden

![Diagram showing the process of using CNN features for recognition]

[Bar chart comparing various performance metrics]
Image Captioning

*man in black shirt is playing guitar.*

*construction worker in orange safety vest is working on road.*

*two young girls are playing with lego toy.*

*boy is doing backflip on wakeboard.*

*girl in pink dress is jumping in air.*

*black and white dog jumps over bar.*

*young girl in pink shirt is swimming in swimming pool.*

*man in blue wetsuit is surfing on wave.*
Material Segmentation [CVPR15]
CNNs + CRFs

CRF Runtime: \(~1\text{s for 640x480 image}\)

\[
E(\mathbf{x} | \mathbf{I}, \theta) = \sum_{i} \psi_i(x_i | \theta) + \sum_{i<j} \psi_{ij}(x_i, x_j | \theta)
\]
### ConvNets breakthroughs for visual tasks

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Performance</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Sermanet et al 2014]: OverFeat (fine-tuned features for each task) (tasks are ordered by increasing difficulty)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>image classification</td>
<td>ImageNet LSVRC 2013</td>
<td>competitive</td>
</tr>
<tr>
<td>object localization</td>
<td>Dogs vs Cats Kaggle challenge 2014</td>
<td>state of the art</td>
</tr>
<tr>
<td>object detection</td>
<td>ImageNet LSVRC 2013</td>
<td>state of the art</td>
</tr>
<tr>
<td></td>
<td>ImageNet LSVRC 2013</td>
<td>competitive</td>
</tr>
<tr>
<td>[Razavian et al, 2014]: public OverFeat library (no retraining) + SVM (simplest approach possible on purpose, no attempt at more complex classifiers) (tasks are ordered by “distance” from classification task on which OverFeat was trained)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>image classification</td>
<td>Pascal VOC 2007</td>
<td>competitive</td>
</tr>
<tr>
<td>scene recognition</td>
<td>MIT-67</td>
<td>state of the art</td>
</tr>
<tr>
<td>fine grained recognition</td>
<td>Caltech-UCSD Birds 200-2011</td>
<td>competitive</td>
</tr>
<tr>
<td></td>
<td>Oxford 102 Flowers</td>
<td>state of the art</td>
</tr>
<tr>
<td>attribute detection</td>
<td>UIUC 64 object attributes</td>
<td>state of the art</td>
</tr>
<tr>
<td></td>
<td>H3D Human Attributes</td>
<td>competitive</td>
</tr>
<tr>
<td>image retrieval (search by image similarity)</td>
<td>Oxford 5k buildings</td>
<td>state of the art</td>
</tr>
<tr>
<td></td>
<td>Paris 6k buildings</td>
<td>state of the art</td>
</tr>
<tr>
<td></td>
<td>Sculp6k</td>
<td>competitive</td>
</tr>
<tr>
<td></td>
<td>Holidays</td>
<td>state of the art</td>
</tr>
<tr>
<td></td>
<td>UKBench</td>
<td>state of the art</td>
</tr>
</tbody>
</table>

## ConvNets breakthroughs for visual tasks

<table>
<thead>
<tr>
<th>[Zeiler et al 2013]</th>
<th>Dataset</th>
<th>Performance</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>image classification</td>
<td>ImageNet LSVRC 2013</td>
<td>state of the art</td>
<td>11.2% error</td>
</tr>
<tr>
<td>Caltech-101 (15, 30 samples per class)</td>
<td>competitive</td>
<td>83.8%, 86.5%</td>
<td></td>
</tr>
<tr>
<td>Caltech-256 (15, 60 samples per class)</td>
<td>state of the art</td>
<td>65.7%, 74.2%</td>
<td></td>
</tr>
<tr>
<td>Pascal VOC 2012</td>
<td>competitive</td>
<td>79% mAP</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>[Donahue et al, 2014]: DeCAF+SVM</th>
<th>Dataset</th>
<th>Performance</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>image classification</td>
<td>Caltech-101 (30 classes)</td>
<td>state of the art</td>
<td>86.91%</td>
</tr>
<tr>
<td>domain adaptation</td>
<td>Amazon -&gt; Webcam, DSLR -&gt; Webcam</td>
<td>state of the art</td>
<td>82.1%, 94.8%</td>
</tr>
<tr>
<td>fine grained recognition</td>
<td>Caltech-UCSD Birds 200-2011</td>
<td>state of the art</td>
<td>65.0%</td>
</tr>
<tr>
<td>scene recognition</td>
<td>SUN-397</td>
<td>competitive</td>
<td>40.9%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>[Girshick et al, 2013]</th>
<th>Dataset</th>
<th>Performance</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>image detection</td>
<td>Pascal VOC 2007</td>
<td>state of the art</td>
<td>48.0% mAP</td>
</tr>
<tr>
<td>Pascal VOC 2010 (comp4)</td>
<td>state of the art</td>
<td>43.5% mAP</td>
<td></td>
</tr>
<tr>
<td>ImageNet LSVRC 2013</td>
<td>state of the art</td>
<td>31.4% mAP</td>
<td></td>
</tr>
<tr>
<td>Pascal VOC 2011 (comp6)</td>
<td>state of the art</td>
<td>47.9% mAP</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>[Oquab et al, 2013]</th>
<th>Dataset</th>
<th>Performance</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>image classification</td>
<td>Pascal VOC 2007</td>
<td>state of the art</td>
<td>77.7% mAP</td>
</tr>
<tr>
<td>Pascal VOC 2012</td>
<td>state of the art</td>
<td>82.8% mAP</td>
<td></td>
</tr>
<tr>
<td>Pascal VOC 2012 (action classification)</td>
<td>state of the art</td>
<td>70.2% mAP</td>
<td></td>
</tr>
</tbody>
</table>

---


M. Oquab, L. Bottou, I. Laptev, and J. Sivic. *Learning and transferring mid-level image representations using convolutional neural networks*. Technical Report HAL-00911179, INRIA, 2013, [http://hal.inria.fr/hal-00911179](http://hal.inria.fr/hal-00911179)
### ConvNets breakthroughs for visual tasks

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Performance</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>[Khan et al 2014]</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• shadow detection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UCF</td>
<td>state of the art</td>
<td>90.56%</td>
</tr>
<tr>
<td>CMU</td>
<td>state of the art</td>
<td>88.79%</td>
</tr>
<tr>
<td>UIUC</td>
<td>state of the art</td>
<td>93.16%</td>
</tr>
<tr>
<td><strong>[Sander Dieleman, 2014]</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• image attributes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kaggle Galaxy Zoo challenge</td>
<td>state of the art</td>
<td>0.07492</td>
</tr>
</tbody>
</table>

“It can be concluded that from now on, deep learning with CNN has to be considered as the primary candidate in essentially any visual recognition task.”

[Razavian 2014]
CNNs at Google (as of 2014)
CNNs at Google (as of 2014)

Applications - Photo Search

- my photos of coffee
- my photos of cake
- my photos of waterfalls
CNNs at Google (as of 2014)

Google Photos - Auto Awesome
CNNs at Google (as of 2014)

More Image Understanding at Google

YouTube
Google Shopping
Advertising
Much more...

StreetView / Maps
Self-Driving Cars
Robotics
CNNs at Google (as of 2014)

The Deep and now Deeper Hammer

Deep learning infrastructure by the Google Brain team

“ImageNet Classification with Deep Convolutional Neural Networks”, Krizhevsky, Sutskever, Hinton, NIPS 2012
CNNs at Google (as of 2014)

Personal Photos - Example Annotations

Christmas tree
Red
Christmas decoration
Christmas

Crowd
Cheering
People
Stadium

Play
Meal
Cake
Child

Hummingbird
Macro photography
Reflection
Red