Large Scale Visual Recognition Challenge (ILSVRC) 2013: Object Detection

Olga Russakovsky (Stanford U.)  Jia Deng (U. of Michigan)  Jonathan Krause (Stanford U.)  Alexander Berg (UNC Chapel Hill)  Fei-Fei Li (Stanford U.)
Agenda

8:30 Classification & localization

8:50 Clarifai
9:05 NUS
9:20 Oxford
9:35 NYU
9:50 Spotlights

10:30 Detection

10:50 NEC Laboratories America

11:00 University of Amsterdam
11:30 Spotlights

Noon Discussion panel

14:00 Invited talk by Vittorio Ferrari:
   Auto-annotation and self-assessment in ImageNet

14:40 Fine-Grained Challenge 2013

http://www.image-net.org/challenges/LSVRC/2013/iccv2013
ILSVRC Task 1: Detection

Fully annotated 200 object classes across 60,000 images

Allows evaluation of generic object detection in cluttered scenes at scale

Modeled after PASCAL VOC
ILSVRC 2013 detection data collection

Step 1: Define object classes

1000 ILSVRC 2012 object classes
ILSVRC 2013 detection data collection

Step 1: Define object classes

1000 ILSVRC 2012 object classes

Discard if "big" in images

T-shirt
ILSVRC 2013 detection data collection

Step 1: Define object classes

- 1000 ILSVRC 2012 object classes
- Discard if “big” in images
- Discard if not “well-suited” for detection

Examples:
- T-shirt
- Hay
ILSVRC 2013 detection data collection

Step 1: Define object classes

1000 ILSVRC 2012 object classes

Discard if “big” in images

Discard if not “well-suited” for detection

494 object classes

T-shirt

Hay
ILSVRC 2013 detection data collection

Step 1: Define object classes

1000 ILSVRC 2012 object classes

Discard if “big” in images

494 object classes

Discard if not “well-suited” for detection

Merge into basic-level classes

200 ILSVRC 2013 DET object classes
ILSVRC 2013 detection data collection

Step 1: Define object classes

Step 2: Collect images

**Train** = **Pos**: ILSVRC 2012 train images for the 200 object classes

- 417 – 67K per class (median 561)

**Neg**: additional images, mostly from Flickr

- 185-10K per class (median 4130)
ILSVRC 2013 detection data collection

Step 1: Define object classes

Step 2: Collect images

**Train** = **Pos**: ILSVRC 2012 train images for the 200 object classes

417 – 67K per class (median 561)

**Neg**: additional images, mostly from Flickr

185-10K per class (median 4130)

**Val, Test** = ILSVRC 2012 val, test for the 200 object classes

Images with target object occupying ≥ 50% of image area

77% (15,522 val and 30,901 test)
ILSVRC 2013 detection data collection

Step 1: Define object classes

Step 2: Collect images

Train

$\begin{align*}
\text{Pos: ILSVRC 2012 train images} & \quad \text{for the 200 object classes} \\
417 \text{ – 67K per class (median 561)} & + \\
\text{Neg: additional images, mostly from Flickr} & 185-10K per class (median 4130)
\end{align*}$

Val, Test

$\begin{align*}
\text{ILSVRC 2012 val, test for} & \quad \text{the 200 object classes} \\
\text{Images with target object occupying} & \quad \geq 50\% \text{ of image area} \\
77\% (15,522 \text{ val and 30,901 test}) & + \\
\text{Additional images from Flickr (queries e.g., "kitchenette," "Australian zoo")} & 23\% (4,599 \text{ val and 9,251 test})
\end{align*}$
ILSVRC 2013 detection data collection

Step 1: Define object classes
Step 2: Collect images
Step 3: Annotate validation&test images completely with all object classes
ILSVRC 2013 detection data collection

Step 1: Define object classes
Step 2: Collect images

Step 3: Annotate validation & test images completely with all object classes

(a) Hierarchical image annotation
ILSVRC 2013 detection data collection

Step 1: Define object classes
Step 2: Collect images
Step 3: Annotate validation & test images completely with all object classes

(a) Hierarchical image annotation

(b) Bounding box annotation
ILSVRC 2013 detection data
ILSVRC 2013 detection data

<table>
<thead>
<tr>
<th>Statistics</th>
<th>PASCAL VOC 2012</th>
<th>ILSVRC 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object classes</td>
<td>20</td>
<td>10x 200</td>
</tr>
<tr>
<td>Training</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Images</td>
<td>5.7K</td>
<td>395K</td>
</tr>
<tr>
<td>Objects</td>
<td>13.6K</td>
<td>25x 345K</td>
</tr>
<tr>
<td>Validation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Images</td>
<td>5.8K</td>
<td>20.1K</td>
</tr>
<tr>
<td>Objects</td>
<td>13.8K</td>
<td>4x 55.5K</td>
</tr>
<tr>
<td>Testing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Images</td>
<td>11.0K</td>
<td>4x 40.1K</td>
</tr>
<tr>
<td>Objects</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>
# ILSVRC 2013 people detection

<table>
<thead>
<tr>
<th>Statistics</th>
<th>ILSVRC 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train</strong></td>
<td></td>
</tr>
<tr>
<td>Positive images</td>
<td>9,877</td>
</tr>
<tr>
<td>Instances</td>
<td>17,728</td>
</tr>
<tr>
<td>Negative images</td>
<td>2,248</td>
</tr>
<tr>
<td><strong>Validation</strong></td>
<td></td>
</tr>
<tr>
<td>Positive images</td>
<td>5,756</td>
</tr>
<tr>
<td>Instances</td>
<td>12,824</td>
</tr>
<tr>
<td>Negative images</td>
<td>14,365</td>
</tr>
</tbody>
</table>

*More than 50,000 person instances annotated (train+val+test)*


## ILSVRC 2013 detection data

<table>
<thead>
<tr>
<th>Property</th>
<th>PASCAL VOC 2012</th>
<th>ILSVRC 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average image resolution</td>
<td>469 x 387 pixels</td>
<td>482 x 415 pixels</td>
</tr>
<tr>
<td>Average object classes per image</td>
<td>1.521</td>
<td>1.534</td>
</tr>
<tr>
<td>Average object instances per image</td>
<td>2.711</td>
<td>2.758</td>
</tr>
<tr>
<td>Average object scale (per instance)</td>
<td>0.207</td>
<td>0.170</td>
</tr>
</tbody>
</table>

Average object scale (per class) computed on validation sets

- Min: 0.01
- 1st quartile: 0.09
- Median: 0.15
- 3rd quartile: 0.23
- Max: 0.44
ILSVRC 2013 detection evaluation

Evaluation modeled after PASCAL VOC:

- Algorithm outputs a list of bounding box detections with confidences
- A detection is considered correct if IOU with ground truth > threshold
- Evaluated by average precision per object class
- Winners of challenge is the team that wins the most object categories

Threshold for Correct Detection

\[
\text{Intersection} \geq 0.5
\]

\[
\text{Union}
\]
Threshold for Correct Detection

\[
\frac{\text{Intersection}}{\text{Union}} \geq 0.5
\]
Threshold for Correct Detection

\[
\frac{\text{Intersection}}{\text{Union}} \geq \min(0.5, \quad )
\]

If within \(\sim 5\) pixels

- Ground truth
- Good detection
- Bad detection
Threshold for Correct Detection

\[
\frac{\text{Intersection}}{\text{Union}} \geq \min(0.5, \frac{w_{gt} h_{gt}}{(w_{gt}+10) (h_{gt}+10)})
\]

*Only matters for small boxes (< ~25 pixels), which is about 5.5% of cases*
ILSVRC 2013 detection results

<table>
<thead>
<tr>
<th>Team Name</th>
<th>mAP</th>
<th># categories won</th>
</tr>
</thead>
<tbody>
<tr>
<td>UvA-Euvison</td>
<td>0.226</td>
<td>130</td>
</tr>
<tr>
<td>NEC-MU (with outside data)</td>
<td>0.209</td>
<td>---</td>
</tr>
<tr>
<td>NEC-MU</td>
<td>0.196</td>
<td>25+35 (2 entries)</td>
</tr>
<tr>
<td>OverFeat-NYU (with outside data)</td>
<td>0.194</td>
<td>---</td>
</tr>
<tr>
<td>Toronto A</td>
<td>0.115</td>
<td>6+1 (2 entries)</td>
</tr>
<tr>
<td>SYSU_Vision</td>
<td>0.105</td>
<td>3</td>
</tr>
<tr>
<td>GPU_UCLA</td>
<td>0.098</td>
<td>0</td>
</tr>
<tr>
<td>Delta</td>
<td>0.061</td>
<td>0</td>
</tr>
<tr>
<td>UIUC-IFP</td>
<td>0.010</td>
<td>0</td>
</tr>
</tbody>
</table>
## ILSVRC 2013 detection results

<table>
<thead>
<tr>
<th>Team Name</th>
<th>mAP</th>
<th># categories won</th>
</tr>
</thead>
<tbody>
<tr>
<td>UvA-Euvision</td>
<td>0.226</td>
<td>130</td>
</tr>
<tr>
<td>NEC-MU (with outside data)</td>
<td>0.209</td>
<td>---</td>
</tr>
<tr>
<td>NEC-MU</td>
<td>0.196</td>
<td>25+35 (2 entries)</td>
</tr>
<tr>
<td>OverFeat-NYU (with outside data)</td>
<td>0.194</td>
<td>---</td>
</tr>
<tr>
<td>Toronto A</td>
<td>0.115</td>
<td>6+1 (2 entries)</td>
</tr>
<tr>
<td>SYSU_Vision</td>
<td>0.105</td>
<td>3</td>
</tr>
<tr>
<td>GPU_UCLA</td>
<td>0.098</td>
<td>0</td>
</tr>
<tr>
<td>Delta</td>
<td>0.061</td>
<td>0</td>
</tr>
<tr>
<td>UIUC-IFP</td>
<td>0.010</td>
<td>0</td>
</tr>
</tbody>
</table>

- **Winner**: NEC-MU (with outside data)
- **Runner-up**: NEC-MU

---


1. **NEC Labs America**, 2. **University of Missouri**

---

**Team Members**

- **Winner**
  - Koen van de Sande
  - Daniel Frontijne
  - Cees Snoek
  - Harro Stokman
  - Arnold Smeulders
  - University of Amsterdam, Euvision Technologies

- **Runner-up**
  - Xiaoyu Wang
  - Miao Sun
  - Tianbao Yang
  - Yuanqing Lin
  - Tony X. Han
  - Shenghuo Zhu

1. **NEC Labs America**, 2. **University of Missouri**
“Easy” detection classes

<table>
<thead>
<tr>
<th>Object class</th>
<th>Best AP (across all entries)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Butterfly</td>
<td>0.827</td>
</tr>
<tr>
<td>Rabbit</td>
<td>0.719</td>
</tr>
<tr>
<td>Snowplow</td>
<td>0.654</td>
</tr>
<tr>
<td>Frog</td>
<td>0.649</td>
</tr>
<tr>
<td>Red panda</td>
<td>0.646</td>
</tr>
<tr>
<td>Tiger</td>
<td>0.628</td>
</tr>
<tr>
<td>Dog</td>
<td>0.625</td>
</tr>
<tr>
<td>Basketball</td>
<td>0.617</td>
</tr>
<tr>
<td>Volleyball</td>
<td>0.613</td>
</tr>
<tr>
<td>Armadillo</td>
<td>0.604</td>
</tr>
</tbody>
</table>
## “Hard” detection classes

<table>
<thead>
<tr>
<th>Object class</th>
<th>Best AP (across all entries)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ladle</td>
<td>0.011</td>
</tr>
<tr>
<td>Microphone</td>
<td>0.016</td>
</tr>
<tr>
<td>Backpack</td>
<td>0.018</td>
</tr>
<tr>
<td>Nail</td>
<td>0.018</td>
</tr>
<tr>
<td>Spatula</td>
<td>0.019</td>
</tr>
<tr>
<td>Plastic bag</td>
<td>0.032</td>
</tr>
<tr>
<td>Axe</td>
<td>0.035</td>
</tr>
<tr>
<td>Purse</td>
<td>0.038</td>
</tr>
<tr>
<td>Water bottle</td>
<td>0.046</td>
</tr>
<tr>
<td>Syringe</td>
<td>0.046</td>
</tr>
</tbody>
</table>
Each dot corresponds to one object class.
ILSVRC2013 detection

Each dot corresponds to one object class

Average precision (best across all entries)

Object scale (on validation set)
ILSVRC2013 detection

Each dot corresponds to one object class.
Useful resources

CloudCV: Object Detection

http://cloudcv.org/objdetect/

- Cached versions of 14 popular scene descriptors on 1.2 million images in ILSVRC2013-CLSLOC
- Pre-trained DPM models on ILSVRC2013-DET

Decaf / Caffe
a Berkeley Vision Project

http://caffe.berkeleyvision.org

- Fast and modifiable implementation of state-of-the-art deep learning algorithms
- Pre-trained classification models on ISLVRC2013-CLSLOC
ILSVRC2013 detection agenda

10:50 Xiaoyu Wang, *Regionlets for Generic Object Detection*

11:10 Koen van de Sande and Daniel Fontijne, University of Amsterdam and Evision Technologies at ILSVRC 2013

11:30 ILSVRC detection spotlights

11:40 Ross Girshick, *Rich feature hierarchies for accurate object detection and semantic segmentation*