IMAGENET Large Scale Visual Recognition Challenge (ILSVRC) 2015
Object Detection from Video (VID)

Wei Liu
UNC Chapel Hill

Olga Russakovsky
CMU

Jia Deng
Univ. of Michigan

Fei-Fei Li
Stanford

Alex Berg
UNC Chapel Hill
Motivation

Why do we need such task?
Motivation 1: Power of benchmark dataset

Classification:
- 2010: 0.28
- 2011: 0.26
- 2012: 0.16
- 2013: 0.12
- 2014: 0.07
- 2015: 0.036

Localization:
- 2011: 0.43
- 2012: 0.34
- 2013: 0.30
- 2014: 0.25
- 2015: 0.09

Detection:
- 2013: 0.23
- 2014: 0.44
- 2015: 0.62

ILSVRC year

1.9x
2.8x
1.4x
Motivation 2: Video is important

Challenges:
- View point change
- Illumination variation
- Motion blur
- Occlusion
- ...
Motivation 2: Video is important

Different from object tracking

Need manual initialization!

TLD tracker

Need unified fast and good detector/tracker in videos!
Motivation 3: We do not have such dataset

A lot of image datasets

A lot of activity dataset

Video dataset with localization is insufficient

Youtube-Objects Dataset
ILSVRC 2015 taster challenge:
Object detection from video (VID)

Fully annotated 30 object classes across 5,354 snippets

Allows evaluation of generic object detection in cluttered videos at scale
ILSVRC 2015 video data collection

• Step 1: Define object categories

- 200 ILSVRC2013 object categories
  - Discard "static" categories
  - Discard if "too many" for detection
- 48 object categories
  - Discard if "not enough" videos
- 30 ILSVRC2015 VID object categories
ILSVRC 2015 video data collection

• Step 1: Define object categories
• Step 2: Collect snippets

Cat videos

Manually cleanup

Segment video to snippets

Manually cleanup

Tabby cat

Creative Common License

Has cat in video
Fix missing class
No cartoon video
No music video

Manually segment

Not too crowded
Not too simple

Freebase

YouTube Data API

/m/0g4cd0
... /
/m/068m4y

Savannah cat
ILSVRC 2015 video data collection

• Step 1: Define object categories
• Step 2: Collect videos
• Step 3: Annotate bounding boxes completely for all categories

http://web.mit.edu/vondrick/vatic/
ILSVRC 2015 video data collection

http://github.com/weiliu89/vatic/tree/vid
ILSVRC object detection from video (VID) task

Evaluation modeled after PASCAL VOC:

- Algorithm outputs a list of bounding box detections with confidences
- A detection is considered correct if intersection over union (IoU) overlap with ground truth > threshold (0.5)
- Evaluated by average precision per object class
- Winners of challenge is the team that wins the most object categories

## ILSVRC2015 VID results – with “provided” data

<table>
<thead>
<tr>
<th>Team Name</th>
<th>Number of categories won</th>
<th>Mean Average Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUvideo</td>
<td>28</td>
<td>67.8</td>
</tr>
<tr>
<td>RUC_BDAI</td>
<td>2</td>
<td>36.0</td>
</tr>
<tr>
<td>ITLab VID - Inha</td>
<td>0</td>
<td>51.5</td>
</tr>
<tr>
<td>UIUC-IFP</td>
<td>0</td>
<td>48.7</td>
</tr>
<tr>
<td>Trimps-Soushen</td>
<td>0</td>
<td>46.1</td>
</tr>
<tr>
<td>1-HKUST</td>
<td>0</td>
<td>42.1</td>
</tr>
<tr>
<td>HiVision</td>
<td>0</td>
<td>37.5</td>
</tr>
<tr>
<td>NICAL</td>
<td>0</td>
<td>23.0</td>
</tr>
<tr>
<td>FACEALL-BUPT</td>
<td>0</td>
<td>22.2</td>
</tr>
<tr>
<td>MPG_UT</td>
<td>0</td>
<td>12.8</td>
</tr>
</tbody>
</table>

CUvideo: Wanli Ouyang, Kai Kang, Junjie Yan, Xingyu Zeng, Hongsheng Li, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang
The Chinese University of Hong Kong

RUC_BDAI: Peng Han, Wenwu Yuan, Zhiwu Lu, Jirong Wen
Renmin University of China
### ILSVRC2015 VID results – with “provided” data

<table>
<thead>
<tr>
<th>Team Name</th>
<th>Mean Average Precision (%)</th>
<th>Number of categories won</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUvideo</td>
<td>67.8</td>
<td>28</td>
</tr>
<tr>
<td>ITLab VID - Inha</td>
<td>51.5</td>
<td>0</td>
</tr>
<tr>
<td>UIUC-IFP</td>
<td>48.7</td>
<td>0</td>
</tr>
<tr>
<td>Trimps-Soushen</td>
<td>46.1</td>
<td>0</td>
</tr>
<tr>
<td>1-HKUST</td>
<td>42.1</td>
<td>0</td>
</tr>
<tr>
<td>HiVision</td>
<td>37.5</td>
<td>0</td>
</tr>
<tr>
<td>RUC_BDAI</td>
<td>36.0</td>
<td>2</td>
</tr>
<tr>
<td>NICAL</td>
<td>23.0</td>
<td>0</td>
</tr>
<tr>
<td>FACEALL-BUPT</td>
<td>22.2</td>
<td>0</td>
</tr>
<tr>
<td>MPG_UT</td>
<td>12.8</td>
<td>0</td>
</tr>
</tbody>
</table>
ILSVRC2015 VID results – with “external” data

<table>
<thead>
<tr>
<th>Team Name</th>
<th>Number of categories won</th>
<th>Mean Average Precision(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amax</td>
<td>18</td>
<td>73.1</td>
</tr>
<tr>
<td>CUvideo</td>
<td>11</td>
<td>69.7</td>
</tr>
<tr>
<td>Trimps-Soushen</td>
<td>1</td>
<td>48.1</td>
</tr>
<tr>
<td>BAD</td>
<td>0</td>
<td>43.4</td>
</tr>
<tr>
<td>NECLAMA</td>
<td>0</td>
<td>32.6</td>
</tr>
</tbody>
</table>

Amax: Jiankang Deng(1,2), Jing Yang(2), Shaoli Huang(1), Hui Shuai(2), Yi Wu(2), Qingshan Liu(2), Dacheng Tao(1)
(1)University of Technology, Sydney
(2)Nanjing University of Information Science & Technology

CUvideo: Wanli Ouyang, Kai Kang , Junjie Yan, Xingyu Zeng, Hongsheng Li, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang
The Chinese University of Hong Kong
ILSVRC2015 VID results – with “external” data

<table>
<thead>
<tr>
<th>Team Name</th>
<th>Mean Average Precision(%)</th>
<th>Number of categories won</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amax</td>
<td>73.1</td>
<td>18</td>
</tr>
<tr>
<td>CUvideo</td>
<td>69.7</td>
<td>11</td>
</tr>
<tr>
<td>Trimps-Soushen</td>
<td>48.7</td>
<td>0</td>
</tr>
<tr>
<td>BAD</td>
<td>43.4</td>
<td>0</td>
</tr>
<tr>
<td>NECLAMA</td>
<td>32.6</td>
<td>0</td>
</tr>
</tbody>
</table>
Object detection from video (VID)

Winner with “provided” data

CUVideo:
Wanli Ouyang, Kai Kang, Junjie Yan, Xingyu Zeng, Hongsheng Li, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang
The Chinese University of Hong Kong

Winner with “external” data

Amax:
Jiankang Deng(1,2), Jing Yang(2), Shaoli Huang(1), Hui Shuai(2), Yi Wu(2), Qingshan Liu(2), Dacheng Tao(1)
(1)University of Technology, Sydney
(2)Nanjing University of Information Science & Technology