Large Scale Visual Recognition Challenge (ILSVRC) 2016
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
Result in ILSVRC VID over the years

Video Detection

<table>
<thead>
<tr>
<th>Year</th>
<th>Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>0.68</td>
</tr>
<tr>
<td>2016</td>
<td>0.81</td>
</tr>
</tbody>
</table>
Object detection from video (VID)

Fully annotated 30 object classes across 6,278 snippets (train+test)

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

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

Cat videos

Manually cleanup
Segment video to snippets
Manually cleanup

Creative Common License
Has cat in video
No cartoon video
No music video
Shot detection
Manually segment
Not too crowded
Not too simple

Tabby cat
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/
video data collection

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

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

ILSVRC object detection from video (VID)

New metric: Take tracking into account

- Algorithm outputs a list of bounding box detections with confidences and tracklet ID.
- Tracklets are sorted by the mean confidence.
- A tracklet is considered correct if intersection over union (IoU) overlap with ground truth tracklet > threshold (0.25, 0.5, 0.75).
- Evaluated by average precision per class. Final score is an average over different thresholds.
- Winners of challenge is the team that has the highest score.
New metric: Take tracking into account

Exemplar video with ground truth
New metric: Take tracking into account

Exemplar video with detection results
Step 1: Compute mean score for tracklet

Exemplar video with detection results
Step 2: Sort tracklets based on mean score

Exemplar video with detection results
Step 3: Evaluate each tracklet

1. Every frame uses bbox IoU overlap 0.5 to determine if it is a true positive detection or not.
2. Then, compute the IoU overlap between detected tracklet and ground truth tracklet.

\[
\text{tracklet IoU} = \frac{\text{# of detected frames}}{\text{# of union frames}} = \frac{3}{4} = 0.75 > 0.5 \text{ (a threshold)}
\]

Exemplar video with detection results
Step 3: Evaluate each tracklet

Exemplar video with detection results

\[
\text{tracklet IoU} = \frac{\# \text{ of detected frames}}{\# \text{ of union frames}} = \frac{2}{3} = 0.66
\]
Step 3: Evaluate each tracklet

Exemplar video with detection results

**negative tracklet**

\[
\text{tracklet IoU} = \frac{\text{# of detected frames}}{\text{# of union frames}} = \frac{1}{3} = 0.33
\]

Note: Duplicate detected tracklets are considered as negative as well
## ILSVRC2016 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>NUIST</td>
<td>10</td>
<td>80.8</td>
</tr>
<tr>
<td>CUVideo</td>
<td>9</td>
<td>76.8</td>
</tr>
<tr>
<td>Trimps-Soushen</td>
<td>1</td>
<td>71.0</td>
</tr>
<tr>
<td>MCG-ICT-CAS</td>
<td>0</td>
<td>73.3</td>
</tr>
<tr>
<td>KAIST-SLSP</td>
<td>0</td>
<td>64.3</td>
</tr>
</tbody>
</table>

**NUIST:** Jing Yang, Hui Shuai, Zhengbo Yu, Rongrong Fan, Qiang Ma, Qingshan Liu, Jiankang Deng

**CUvideo:** Hongsheng Li*, Kai Kang*, Wanli Ouyang, Junjie Yan, Tong Xiao, Xingyu Zeng, Kun Wang, Xihui Liu, Qi Chu, Junming Fan, Yucong Zhou, Yu Liu, Ruohui Wang, Shengen Yan, Dahua Lin, Xiaogang Wang

(* indicates equal contribution)

The Chinese University of Hong Kong, SenseTime Group Limited
# ILSVRC2016 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>NUIST</td>
<td>17</td>
<td>79.6</td>
</tr>
<tr>
<td>Trimps-Soushen</td>
<td>5</td>
<td>72.1</td>
</tr>
<tr>
<td>ITLab-Inha</td>
<td>3</td>
<td>73.1</td>
</tr>
<tr>
<td>DPAI Vision</td>
<td>0</td>
<td>61.5</td>
</tr>
<tr>
<td>TEAM1</td>
<td>0</td>
<td>21.8</td>
</tr>
</tbody>
</table>

NUIST: Jing Yang, Hui Shuai, Zhengbo Yu, Rongrong Fan, Qiang Ma, Qingshan Liu, Jiankang Deng

Trimps-Soushen: Jie Shao, Xiaoteng Zhang, Zhengyan Ding, Yixin Zhao, Yanjun Chen, Jianying Zhou, Wenfei Wang, Lin Mei, Chuanping Hu
The Third Research Institute of the Ministry of Public Security, P.R. China.
<table>
<thead>
<tr>
<th>Team Name</th>
<th>Mean Average Precision(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUVideo</td>
<td>55.9</td>
</tr>
<tr>
<td>NUIST</td>
<td>54.9</td>
</tr>
<tr>
<td>MCG-ICT-CAS</td>
<td>48.9</td>
</tr>
<tr>
<td>KAIST-SLSP</td>
<td>32.7</td>
</tr>
<tr>
<td>CIGIT_Media</td>
<td>23.0</td>
</tr>
</tbody>
</table>

CUVideo: Hongsheng Li*, Kai Kang*, Wanli Ouyang, Junjie Yan, Tong Xiao, Xingyu Zeng, Kun Wang, Xihui Liu, Qi Chu, Junming Fan, Yucong Zhou, Yu Liu, Ruohui Wang, Shengen Yan, Dahua Lin, Xiaogang Wang

(* indicates equal contribution)
The Chinese University of Hong Kong, SenseTime Group Limited

NUIST: Jing Yang, Hui Shuai, Zhengbo Yu, Rongrong Fan, Qiang Ma, Qingshan Liu, Jiankang Deng
# ILSVRC2016 VID Tracking Results – with “external” data

<table>
<thead>
<tr>
<th>Team Name</th>
<th>Mean Average Precision(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUIST</td>
<td>58.4</td>
</tr>
<tr>
<td>ITLab-Inha</td>
<td>49.1</td>
</tr>
</tbody>
</table>

**NUIST:** Jing Yang, Hui Shuai, Zhengbo Yu, Rongrong Fan, Qiang Ma, Qingshan Liu, Jiankang Deng

**ITLab-Inha:** Byungjae Lee\(^1\), Songguo Jin\(^1\), Enkhbayar Erdenee\(^1\), Mi Young Nam\(^2\), Young Giu Jung\(^2\), Phill Kyu Rhee\(^1\)

1. Inha University
2. NaeulTech
Object detection from video (VID)

Winner with “provided” and “external” data

NUIST:
Jing Yang, Hui Shuai,
Zhengbo Yu, Rongrong
Fan, Qiang Ma, Qingshan
Liu, Jiankang Deng
Object detection from video with tracking

Winner with “provided” data
CUvideo:
Hongsheng Li*, Kai Kang*, Wanli Ouyang, Junjie Yan, Tong Xiao, Xingyu Zeng, Kun Wang, Xihui Liu, Qi Chu, Junming Fan, Yucong Zhou, Yu Liu, Ruohui Wang, Shengen Yan, Dahua Lin, Xiaogang Wang (* indicates equal contribution)
The Chinese University of Hong Kong,
SenseTime Group Limited

Winner with “external” data
NUIST:
Jing Yang, Hui Shuai, Zhengbo Yu, Rongrong Fan, Qiang Ma, Qingshan Liu, Jiankang Deng