Trimps at ILSVRC2015

Jie SHAO, Xiaoteng ZHANG, Jianying ZHOU, Zhengyan DING, Wenfei WANG, Lin MEI, Chuanping HU

The Third Research Institute of the Ministry of Public Security, P.R. China.

jieshao.mail@gmail.com

17 December 2015
Summary of Trimps Submission

- **Object localization**
  - 2\textsuperscript{nd} place, 12.29\% error (1\textsuperscript{st} place with extra data)

- **Object detection from video (VID)**
  - 4\textsuperscript{th} place, 0.461 mAP (3\textsuperscript{rd} place with extra data)

- **Scene classification**
  - 4\textsuperscript{th} place, 17.98\% error

- **Object detection**
  - 7\textsuperscript{th} place, 0.446 mAP (4\textsuperscript{th} place with extra data)
Object Localization

• Simple pipeline

Input Image ➔ Classification ➔ Top-5 Labels ➔ Localization

- Label-1 ➔ Box-1
- Label-2 ➔ Box-2
- Label-3 ➔ Box-3
- Label-4 ➔ Box-4
- Label-5 ➔ Box-5
Object Localization — CLS

• Training
  – Multiple CNN models with large diversity
    • 7 * BN-Inception (32 Layers)
    • 2 * MSRA-Net (22 Layers)

  – Data augmentation
    • Random crops, multi-scale, contrast and color jittering
Object Localization — CLS

• **Testing for single model**
  – Multi-scale densely crop
  – Overfeat-style augmentation
Object Localization — CLS

• Testing for multi-model
  – Scores Fusion (+1.07% accuracy)
    \[ S = \sum_{i=1}^{N} W_i \times Score_i, \quad Labels = f_{top5}(S) \]
  – Labels Fusion (+1.17% accuracy)
    • Keep M labels for single model, N models got N*M labels
    • Select 5 most labels from N*M
Object Localization — CLS

• Top-5 classification error (test set)
Object Localization — LOC

• Based on Fast R-CNN
  – Pre-trained models: VGG16, VGG19, GoogLeNet
  – Region proposals: EdgeBoxes + Filter(\sim 500/\text{img})
Object Localization — LOC

• **Single model improvements**
  – Objectness loss
  – Negative categories
  – Bounding box voting

• **Ensemble**

![Bar chart showing Top-5 error comparison for different models on the Val set.](chart)

- **Baseline**: 14.25
- **Improved**: 13.58
- **Ensemble**: 12.29
Object Localization — LOC

• Fast R-CNN

Fast R-CNN, Girshick R. 2015
Object Localization — LOC

• Negative categories and objectness loss
Object Localization — LOC

• Negative categories (training)
  – Positive: $\text{IOU} \geq 0.5$, Negative: $0.2 \leq \text{IOU} < 0.5$,
  Background: others
Object Localization — LOC

- **Bounding box voting (testing)**

For each category

- Select region $b$ with highest score
- Select regions $R_i$, s.t.

$$IOU(b, R_i) \geq 0.5 \text{ and } score_{R_i} \geq th$$

- Voting using $R+b$, $Box = \frac{\sum_{i=1}^{k} score_i * bbox_i}{\sum_{i=1}^{k} score_i}$

Object detection via a multi-region & semantic segmentation-aware CNN model, Gidaris S, Komodakis N. 2015
Object Localization — LOC

• Multi-model ensemble (testing)
  – Bounding box voting (+0.3% vs best single model)
  – Most crowded (not highest scored, +1.4%)
Object Localization — LOC

- Top-5 localization error (test set)

Object Localization (rank #2)
Scene Classification

• **Dataset**
  
  – 8.1M train images, unbalanced
  
  – Larger image size, min dimension is 512
  
  – Both background and foreground are important
Scene Classification

• Design
  – Data sweeping
  – Larger input size, deeper and wider network
  – Multi-branch: whole image and part
Scene Classification

• Data sweeping
  – Random sweep training data at each epoch
  – Speed up training without accuracy decline

\[ s(n) = \begin{cases} 
\cos(\lambda n), & n \in [0, l] \\
c, & n \in [l + 1, K] 
\end{cases} \]

Scene Classification

• Larger inception

270x270 → 135x135 → 67x67 → 33x33 → 17x17 → 8x8
Scene Classification

• Two-branch inception
Scene Classification

• Top-5 error (test set)
Object Detection

• Pre-train model
  – VGG16, VGG19, pooling replaced with conv
  – COCO data used in some models

• Negative categories
  – Most improved on val set: +3.2% mAP

• Objectness
  – Most improved on val set: +2.2% mAP

• Bounding box voting
Object Detection

• Results

Object Detection (rank #7)

mAP

2013-UvA 2014-GoogLeNet 2015-MSRA* 2015-Trimps*

0.226 0.439 0.621 0.446

* Larger test set this year
Object Detection from Video

• From 200 to 30
  – Using models from object detection task
  – Using video data to do fine tuning
Object Detection from Video

- Results

Object Detection from Video (rank #4)

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015-CUvideo</td>
<td>0.678</td>
</tr>
<tr>
<td>2015-ITLab VID - Inha</td>
<td>0.515</td>
</tr>
<tr>
<td>2015-UIUC-IFP</td>
<td>0.487</td>
</tr>
<tr>
<td>2015-Trimps</td>
<td>0.461</td>
</tr>
<tr>
<td>2015-1-HKUST</td>
<td>0.421</td>
</tr>
</tbody>
</table>
Acknowledgement

• Professor
  Xiangyang Xue@Fudan

• Professor
  Zheng Zhang@NYU-Shanghai

• Professor
  Xiang Bai@HUST