Good Practices for Deep Feature Fusion

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, China

09 October 2016
Trimps-Soushen@ILSVRC2016

- **Object Localization**
  - 1st place, 7.71% error

- **Object Classification**
  - 1st place, 2.99% error

- **Object Detection**
  - 3rd place, 0.618 mAP

- **Scene Classification**
  - 3rd place, 10.30% error

- **Object Detection from video**
  - 3rd place, 0.71 mAP
Object Localization-CLS

• Different kinds of deep models

Cls Errors for Top-10 Difficult Categories

Object Localization-CLS

• Details
  – Training
    – Multi-scale augmentation & Large mini-batch size
    – Identity map (Pre-activation)
  – Testing
    – Multi-scale & flip & dense fusion

<table>
<thead>
<tr>
<th></th>
<th>Inception-v3</th>
<th>Inception-v4</th>
<th>Inception-Resnet-v2</th>
<th>Resnet-200</th>
<th>Wrn-68-3</th>
<th>Fusion (Val.)</th>
<th>Fusion (Test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Err. (%)</td>
<td>4.20</td>
<td>4.01</td>
<td>3.52</td>
<td>4.26</td>
<td>4.65</td>
<td>2.92 (-0.6)</td>
<td>2.99</td>
</tr>
</tbody>
</table>
Object Localization-CLS

• Fusion Error Analysis
  – Top-k Accuracy on Val. Dataset

✓ Top-20 Accuracy reaches 99%

![Graph showing object classification top-k accuracy]
Object Localization-CLS

• Fusion Error Analysis
  – Manually analyze 1458 error images from Val set
  ✓ Classification Accuracy is hard to improve (1%)

<table>
<thead>
<tr>
<th>Error Categories</th>
<th>Numbers</th>
<th>Percentages(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label May Wrong</td>
<td>221</td>
<td>15.16</td>
</tr>
<tr>
<td>Multiple Objects (&gt;5)</td>
<td>118</td>
<td>8.09</td>
</tr>
<tr>
<td>Non-Obvious Main Object</td>
<td>355</td>
<td>24.35</td>
</tr>
<tr>
<td>Confusing Label</td>
<td>206</td>
<td>14.13</td>
</tr>
<tr>
<td>Fine-grained Label</td>
<td>258</td>
<td>17.70</td>
</tr>
<tr>
<td>Obvious Wrong</td>
<td>234</td>
<td>16.05</td>
</tr>
<tr>
<td>Partial Object</td>
<td>66</td>
<td>4.53</td>
</tr>
</tbody>
</table>
Object Localization-CLS

• Image Examples (Label may wrong)

Predict:
1 pencil box
2 diaper
3 bib
4 purse
5 running shoe

Ground Truth: sleeping bag
Object Localization-CLS

• Image Examples (Multiple objects)

Predict:
1 lion
2 web site
3 frying pan
4 teddy
5 pop bottle

Ground Truth:
icce cream
Object Localization-CLS

• Image Examples (Non-obvious main object)

Predict:
1 dock
2 submarine
3 boathouse
4 breakwater
5 lifeboat

Ground Truth: paper towel
Object Localization-CLS

• Image Examples (Confusing label)

Predict:
1 carton
2 packet
3 toilet tissue
4 vending machine
5 crate

Ground Truth: sunscreen
Object Localization-CLS

• Image Examples (Fine-grained label)

Predict:
1 bolete
2 earthstar
3 gyromitra
4 hen of the woods
5 mushroom

Ground Truth: stinkhorn
Object Localization-CLS

• Image Examples (Obvious wrong)

Predict:
1 apron
2 plastic bag
3 sleeping bag
4 umbrella
5 bulletproof vest

Ground Truth:
poncho
Object Localization-CLS

• Image Examples (Partial object)

Predict:
1 plate
2 carbonara
3 chocolate sauce
4 bakery
5 ice cream

Ground Truth:
restaurant
Object Localization-CLS

• Fusion Error Analysis
  – Manually analyze 1458 error images from Val set
  ✔ Classification Accuracy is hard to improve (1%)

<table>
<thead>
<tr>
<th>Error Categories</th>
<th>Numbers</th>
<th>Percentages(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label May Wrong</td>
<td>221</td>
<td>15.16</td>
</tr>
<tr>
<td>Multiple Objects (&gt;5)</td>
<td>118</td>
<td>8.09</td>
</tr>
<tr>
<td>Non-Obvious Main Object</td>
<td>355</td>
<td>24.35</td>
</tr>
<tr>
<td>Confusing Label</td>
<td>206</td>
<td>14.13</td>
</tr>
<tr>
<td>Fine-grained Label</td>
<td>258</td>
<td>17.70</td>
</tr>
<tr>
<td>Obvious Wrong</td>
<td>234</td>
<td>16.05</td>
</tr>
<tr>
<td>Partial Object</td>
<td>66</td>
<td>4.53</td>
</tr>
</tbody>
</table>
Object Classification

- 2011-XRCE: 25.81%
- 2012-Supervision: 16.42%
- 2013-Clarifai: 11.74%
- 2014-GoLeNet: 6.66%
- 2015-MSRA: 3.56%
- 2016-ResNeXt: 3.03%
- 2016-TrimpS-Soushen: 2.99%
Object Localization-LOC

• Based on Faster R-CNN Style Pipeline


Object Localization-LOC

• Improvements
  – Cascade Fast RCNN(multi)
  – Learn to Rank(single)
  – Fusion in single model(single)
  – Fusion between models(multi)

<table>
<thead>
<tr>
<th>Localization</th>
<th>Val-Top-5 err (%)</th>
<th>Test-Top-5 err (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Single Model</td>
<td>10.35</td>
<td>/</td>
</tr>
<tr>
<td>Single Model Improvements</td>
<td>8.51(-1.84)</td>
<td>/</td>
</tr>
<tr>
<td>Ensemble</td>
<td>7.58(-0.93)</td>
<td>7.71</td>
</tr>
</tbody>
</table>
Object Localization-LOC

• Cascade Fast RCNN *(Region Fusion)*
Object Localization-LOC

• **Learn to Rank (test)**

  For the given category (used in 45 categories)

  – Train new CNN model to rescore regions

  – Average the original score and new score

  – Select regions by the average score

  n04019541
  n04228054
  n02825657
  n03355925
  n09256479
  n03825788
  n09288635
  n04264628
  n03961711
  ......
Object Localization-LOC

• Fusion in single model
  – Boost about 0.4%

\[
\begin{align*}
\text{score} > \text{th} & \quad \rightarrow \quad (x_1, y_1, x_2, y_2, \text{score})_{j=0} \\
\text{score} < \text{th} & \quad \rightarrow \quad (x_1, y_1, x_2, y_2, \text{score})_{j=N}
\end{align*}
\]

\[
\text{Fusion}_\text{box} = \frac{\sum_{i=0}^{K} \text{score}_i \times \text{box}_i}{\sum_{i=0}^{K} \text{score}_i}
\]

• Fusion between models
  – Voting by region cluster
  – Boost more than 0.9%
# Object Localization-LOC

<table>
<thead>
<tr>
<th>Localization</th>
<th>Val-Top-5 err (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>*ResNet-152</td>
<td>8.51</td>
</tr>
<tr>
<td>ResNet-101</td>
<td>8.65</td>
</tr>
<tr>
<td>*Inception-ResNet-V2</td>
<td>8.81</td>
</tr>
<tr>
<td>Inception-V4</td>
<td>8.88</td>
</tr>
<tr>
<td>Inception-V3</td>
<td>9.27</td>
</tr>
<tr>
<td>*WRN-68</td>
<td>8.87</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Localization</th>
<th>Val-Top-5 err (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble all</td>
<td>7.58</td>
</tr>
<tr>
<td>No ResNet-152</td>
<td>7.93(+0.35)</td>
</tr>
<tr>
<td>No Inception-ResNet-V2</td>
<td>7.88(+0.30)</td>
</tr>
<tr>
<td>No WRN-68</td>
<td>7.75(+0.17)</td>
</tr>
</tbody>
</table>

*Diversity between models is important*
Object Localization-LOC

• Details in Localization
  ✓ No weight-decay, no dropout
  ✓ Enough epochs
  ✓ Suitable mini-batch/iter_size, like 8 or 16
  ✓ Diversity between models
  ✓ Different anchor size
  ✓ Better regions
  ✓ Multi-scale train/test
Object Localization-LOC

• Improve localization accuracy is difficult
Object Localization-LOC

• Improve localization accuracy is difficult
Object Localization-LOC

Object Localization

<table>
<thead>
<tr>
<th>Year</th>
<th>Method</th>
<th>Top-5 error(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>SuperVision</td>
<td>33.48</td>
</tr>
<tr>
<td>2013</td>
<td>Overfeat</td>
<td>29.88</td>
</tr>
<tr>
<td>2014</td>
<td>VGG</td>
<td>25.32</td>
</tr>
<tr>
<td>2016</td>
<td>NUIST</td>
<td>9.06</td>
</tr>
<tr>
<td></td>
<td>MSRA</td>
<td>9.01</td>
</tr>
<tr>
<td></td>
<td>Hikvision</td>
<td>8.74</td>
</tr>
<tr>
<td></td>
<td>Trimps-Soushen</td>
<td>7.71</td>
</tr>
</tbody>
</table>
Scene Classification

• Details
  – Training
    – Torch (Memory-shared Optimization)
    – Small scale range & large input crop size

  – Testing
    – Improve multi-scale fusion & multi-model fusion
Scene Classification

- Improve Multi-scale Fusion

Single Model

<table>
<thead>
<tr>
<th>Image ($N_1 \times N_1$)</th>
<th>Full Conv. Net</th>
<th>Max Pool</th>
<th>Concat</th>
<th>FC</th>
<th>SoftMax Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image ($N_2 \times N_2$)</td>
<td>Full Conv. Net</td>
<td>Max Pool</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image ($N_3 \times N_3$)</td>
<td>Full Conv. Net</td>
<td>Max Pool</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results

<table>
<thead>
<tr>
<th></th>
<th>Baseline Model</th>
<th>Single Model</th>
<th>Multi models (Val.)</th>
<th>Selected models (Test.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Err.</td>
<td>12.40%</td>
<td>11.20% (-1.2%)</td>
<td>10.32% (-0.88%)</td>
<td>10.50 (+0.18%)</td>
</tr>
</tbody>
</table>
Scene Classification

- **Improve Multi-model Fusion**
  - Fusion models two by two (*Less Overfitting*)

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image $(N_1 \times N_2)$</td>
<td>Image $(N_1 \times N_3)$</td>
</tr>
<tr>
<td>Full Conv. Net</td>
<td>Full Conv. Net</td>
</tr>
<tr>
<td>Max Pool</td>
<td>Max Pool</td>
</tr>
<tr>
<td>Concat</td>
<td>Concat</td>
</tr>
<tr>
<td>$N_k = N_{k-1} + 32$</td>
<td>$N_k = N_{k-1} + 32$</td>
</tr>
</tbody>
</table>

**Fine-tune on Train. Dataset**

<table>
<thead>
<tr>
<th>Two Models</th>
<th>7*Two Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Val. Err.</td>
<td>10.80 (-0.4%)</td>
</tr>
<tr>
<td>Test. Err.</td>
<td>10.42 (+0.03%)</td>
</tr>
</tbody>
</table>
Scene Classification

Scene Classification

Combine with a Places2-pretrained model

only provided data

Top-5 error (%)

- 2016-NOSCENE: 10.93%
- 2016-NTU-SC: 10.85%
- 2016-SIAT_MMLAB: 10.43%
- 2016-Trimps-Soushen: 10.42%
- 2016-Trimps-Soushen: 10.3%
- 2016-MW: 10.19%
- 2016-Hikvision: 9.01%
Object Detection

• Testing (single: +2~3 mAP; multi: +4.3 mAP)
  – 300 regions: predict boxes B from our best model
  – New 300 regions: new predict boxes using B as input
  – Average softmax and coordinate using 600 regions and their flips across all models
Object Detection

![Object Detection Chart](chart.png)

- 2016-CL: 55.35
- 2016-NUIST: 60.88
- 2016-360+MCG...: 61.56
- 2015-Trimps-Soushen: 61.82
- 2015-MSRA: 62.07
- 2016-Hikvision: 65.27
- 2016-CUimage: 66.28
Object Detection from Video

• From 200 to 30
  – Using models from detection task of 200 classes to do video detection
  – Using video data to do fine-tuning

• Add extra train data
  – Select some train data from ImageNet dataset
  – Using part of Val data to train
<table>
<thead>
<tr>
<th>Dataset</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016-NUS-VISENZE</td>
<td>64.06</td>
</tr>
<tr>
<td>2016-KAIST-SLSP</td>
<td>64.28</td>
</tr>
<tr>
<td>2016-Trimps-Soushen</td>
<td>70.97</td>
</tr>
<tr>
<td>2016-MCG-ICT-CAS</td>
<td>73.31</td>
</tr>
<tr>
<td>2016-CUvideo</td>
<td>76.8</td>
</tr>
<tr>
<td>2016-NUIST</td>
<td>80.83</td>
</tr>
</tbody>
</table>
Thank you!

Contact: jieshao.mail@gmail.com