MSCOCO Keypoints Challenge 2017

Megvii (Face++)
Megvii (Face++)

Team members (Keypoints & Detection):

Yilun Chen*
Zhicheng Wang*
Xiangyu Peng
Zhiqiang Zhang
Gang Yu
Chao Peng
Tete Xiao
Zeming Li
Xiangyu Zhang
Yuning Jiang
Jian Sun
# Results

- **COCO 17 & 16 Keypoints**

|           | AP   | AP50 | AP75 | APm  | APs  | AR   | AR50 | AR75 | ARm  | ARs  |
|-----------|------|------|------|------|------|------|------|------|------|------|------|
| CMU-Pose[1] | 0.605 | 0.834 | 0.664 | 0.551 | 0.681 | 0.659 | 0.864 | 0.713 | 0.594 | 0.748 |
| G-RMI[2]   | 0.598 | 0.81  | 0.651 | 0.567 | 0.667 | 0.664 | 0.865 | 0.712 | 0.618 | 0.726 |
| Ours       | 0.726 | 0.905 | 0.791 | 0.684 | 0.788 | 0.788 | 0.943 | 0.846 | 0.746 | 0.846 |


Note: [1] and [2] are evaluated on COCO 2016 test challenge dataset, while ours method is evaluated on COCO 2017 test challenge dataset.
Overview

- Top-down Pipeline

- Network Design
  - Is Hourglass good for COCO keypoint?
  - Motivation: How human locate keypoints?
  - Our Network Architecture

- Techniques & Experiments

- Conclusion
Overview

- Top-down Pipeline
Top-Down pipeline
Top-Down pipeline
Top-Down pipeline

MegDet → crop → Single Person Pose Estimation Network
Person Detector

- Our person detector is based on MegDet trained on 80-class labeled data, without specific training for person. (Human detection AP is 62.0)

<table>
<thead>
<tr>
<th>Human AP(area = all)</th>
<th>Human AP(area = medium)</th>
<th>Human AP(area = large)</th>
</tr>
</thead>
<tbody>
<tr>
<td>62.0</td>
<td>69.1</td>
<td>78.5</td>
</tr>
</tbody>
</table>
Overview

- Top-down Pipeline
- Network Design
Overview

- Top-down Pipeline
- Network Design
  - Is Hourglass good for COCO keypoint?
Is Hourglass good for COCO keypoint

<table>
<thead>
<tr>
<th>models</th>
<th>input size</th>
<th>FLOPs</th>
<th>param_dim</th>
<th>param_size</th>
<th>depth_conv_fc</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hourglass(^2) 1-stage</td>
<td>256x192</td>
<td>3.9G</td>
<td>3M</td>
<td>12MB</td>
<td>38</td>
<td>0.602</td>
</tr>
<tr>
<td>ResNet-50-FPN(^1)</td>
<td>256x192</td>
<td>3.9G</td>
<td>24M</td>
<td>93MB</td>
<td>51</td>
<td>0.671</td>
</tr>
</tbody>
</table>

- ResNet-FPN-like\(^1\) network works better than hourglass-like\(^2\) network (1-stage) of the same FLOPs.

---

### Is Hourglass good for COCO keypoint localization?

<table>
<thead>
<tr>
<th>Model</th>
<th>FLOPs</th>
<th>Pckh-0.5 (MPI val)</th>
<th>AP@OKS0.75 (COCO val)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-stage hourglass(256*192)</td>
<td>3.9G</td>
<td>0.893</td>
<td>0.663</td>
</tr>
<tr>
<td>2-stage hourglass(256*192)</td>
<td>6.1G</td>
<td>0.921</td>
<td>0.755</td>
</tr>
<tr>
<td>3-stage hourglass(256*192)</td>
<td>8.3G</td>
<td>0.924</td>
<td>0.754</td>
</tr>
<tr>
<td>4-stage hourglass(256*192)</td>
<td>10.5G</td>
<td>0.924</td>
<td>0.752</td>
</tr>
</tbody>
</table>

- Two stages are enough for keypoint localization for better trade-off.
- More stages (stages larger than 2) are not good at high-precision localization, for example @0.75 OKS
  - Guess: Hourglass stages harm the spatial resolution.
Overview

- Top-down Pipeline

- Network Design
  - Is Hourglass good for COCO keypoint?
  - Motivation: How human locates keypoints?
Motivation: How human locate keypoints?

Easy visible parts:
- Nose
- Left elbow
- Right hand

Hard visible parts:
- Left knee
- Right knee
- Left hip

Invisible parts:
- Right shoulder
- Hard to distinguish?
Network’s Design Goal

Input image → Visible easy part → Visible hard part → Invisible part → Output image

- receptive view getting larger & more context
Overview

- Top-down Pipeline

- Network Design
  - Is Hourglass good for COCO keypoint?
  - Motivation: How human locate keypoints?
  - Our Network Architecture
Network Architecture

Network Design Principles:
- Follow the human perspective
  - locate visible easy parts => locate visible hard parts => locate invisible parts
- Two stages
  - VisibleNet: to locate the both the easy parts (earlier layers) and visible hard parts (deep layers)
  - GlobalNet: to locate hard parts as well
Overview

▪ Top-down Pipeline

▪ Network Design
  ▪ Is Hourglass good for COCO keypoint?
  ▪ Motivation: How human locate keypoint?
  ▪ Our Network Architecture

▪ Techniques & Experiments
## Techniques & Experiments

<table>
<thead>
<tr>
<th>Model Configuration</th>
<th>AP% (COCO minival)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (ResNet-50-FPN) (256x192)</td>
<td>67.1</td>
</tr>
<tr>
<td>Our network (ResNet-50) (256x192)</td>
<td>69.0</td>
</tr>
<tr>
<td>Our network (ResNet-50) (384x288)</td>
<td>71.0</td>
</tr>
<tr>
<td>Our network (Inception-ResNet) (384x288)</td>
<td>72.3</td>
</tr>
<tr>
<td>+ Large Batch</td>
<td>73.0</td>
</tr>
</tbody>
</table>

More ablation experiments on our network will come soon in our CVPR submission.
Techniques & Experiments

- Data augmentation (+0.4AP)
  - Crop augmentation
  - Random scales (0.7~1.35)
  - Rotation (-45°~45°)
Techniques & Experiments

- Data augmentation (+0.4AP)
  - Crop augmentation
  - Random scales (0.7~1.35)
  - Rotation (-45º ~ 45º)
- Large Batch (+0.4~0.7AP)
Techniques & Experiments

- Data augmentation (+0.4AP)
  - Crop augmentation
  - Random scales (0.7~ 1.35)
  - Rotation (-45° ~ 45°)
- Large Batch (+0.4~0.7AP)
- Segmentation supervision (+0.2~0.6AP)
  - Enhance the network’s ability to distinguish the detected person from crowded scene.
Techniques & Experiments

- Data augmentation (+0.4AP)
  - Crop augmentation
  - Random scales (0.7~1.35)
  - Rotation (-45°~45°)
- Large Batch (+0.4~0.7AP)
- Segmentation supervision (+0.2~0.6AP)
  - Enhance the network’s ability to distinguish the detected person from crowded scene.
- Ensemble (+1.1~1.5AP)
  - Heatmap merge

<table>
<thead>
<tr>
<th></th>
<th>AP% (COCO minival)</th>
<th>AP% (COCO challenge)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our network with all techniques</td>
<td>74.7</td>
<td>72.6</td>
</tr>
</tbody>
</table>
Illustrative results of our method
Conclusion

- The two-stage network design is crucial.
  - **VisibleNet**: locates both the visible easy parts (earlier layers) and visible hard parts (deep layers)
  - **GlobalNet**: locates invisible parts

- Data augmentation is the key to enhance robustness of network, especially in CNN.

- **Large batch** technique is not only applicable in object detection, but also in keypoint.

- Segmentation supervision is also an universe skill in training CNN.
We are hiring!

@Beijing, @Nanjing, @Seattle

career@megvii.com
Thanks & Questions

wangzhicheng@megvii.com