Encoder-Decoder Based Convolutional Neural Networks with Multi-Scale-Aware Modules for Crowd Counting

Pongpisit Thanasutives¹, K. Fukui¹, M. Numao¹, B Kijsirikul²
¹Osaka University; ²Chulalongkorn University

ICPR 2020

Codes: https://github.com/Pongpisit-Thanasutives/Variations-of-SFANet-for-Crowd-Counting
Introduction (1)

Crowd Counting: To count a number of people in a given image for public safety, surveillance monitoring, etc.

(Image from Zhang et al., 2016 [1])
Introduction (2)

**Problem** (Gao et al., 2019 [2]):
- Heavy occlusion (noisy image, blurred objects)
- Perspective distortion (different camera angles)
- Scale variation (different sizes of head and surrounding context), etc.

**Goal**: Solve these problems using a combination of multi-scale-aware modules and dual-path decoder.
Data preprocessing (based on Zhang et al., 2016 [1]): Convolve the head annotation with Gaussian kernel (G) which has fixed standard deviation (σ). Assuming that there is a head annotation at pixel $x_i$ represented as $\delta(x-x_i)$. The density map $D(x)$ can be defined as

$$D(x) = \sum_{i=1}^{C} \delta(x - x_i) * G_\sigma(x)$$

$A(i) = \begin{cases} 
0 & 0.001 > D(i) \\
1 & 0.001 \leq D(i)
\end{cases}$

A: Attention map
1st Proposed model - M-SFANet (1)

The architecture of M-SFANet

Inspired by SFANet by Zhu et al., 2019 [3].

Atrous spatial pyramid pooling (ASPP) with augmented image-level features by Chen et al., 2018 [4].
1\textsuperscript{st} Proposed model - M-SFANet (2)

Context-aware module (CAN), Liu et al., 2019 [5].

The architecture of M-SFANet
**Proposed model - M-SegNet**

There are no CAN and ASPP to additionally emphasize multi-scale information.

The bilinear upsampling is replaced with max unpooling operation using the memorized max-pooling indices (Badrinarayanan et al., 2017 [6]).

Less computational resources than M-SFANet with competitive performance. More suitable for speed-constrained applications.

The architecture of M-SegNet
## Results (1)

### Performance comparison

<table>
<thead>
<tr>
<th>Method</th>
<th>Part A</th>
<th>Part B</th>
<th>UCF_CC_50</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>RMSE</td>
<td>MAE</td>
</tr>
<tr>
<td>CAN</td>
<td>62.3</td>
<td>100.0</td>
<td>7.8</td>
</tr>
<tr>
<td>SFANet</td>
<td>59.8</td>
<td>99.3</td>
<td>6.9</td>
</tr>
<tr>
<td>S-DCNet</td>
<td>58.3</td>
<td>95.0</td>
<td>6.7</td>
</tr>
<tr>
<td>SANet + SPANet</td>
<td>59.4</td>
<td>92.5</td>
<td>6.5</td>
</tr>
<tr>
<td>M-SegNet</td>
<td>60.55</td>
<td>100.80</td>
<td>6.80</td>
</tr>
<tr>
<td>M-SFANet</td>
<td>59.69</td>
<td>95.66</td>
<td>6.76</td>
</tr>
<tr>
<td>M-SFANet + M-SegNet</td>
<td>57.55</td>
<td>94.48</td>
<td>6.32</td>
</tr>
</tbody>
</table>

More results on the paper.
Results (2)

Ablation study

<table>
<thead>
<tr>
<th>Method</th>
<th>Part A</th>
<th>Part B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>RMSE</td>
</tr>
<tr>
<td>M-SFANet w/o CAN</td>
<td>62.41</td>
<td>101.13</td>
</tr>
<tr>
<td>M-SFANet w/o ASPP</td>
<td>61.25</td>
<td>102.37</td>
</tr>
<tr>
<td>M-SFANet w/o skip conn.</td>
<td><strong>60.07</strong></td>
<td><strong>99.47</strong></td>
</tr>
</tbody>
</table>

ASPP: Suitable for sparse scenes.
CAN: Suitable for dense scenes.
Summary

- For M-SFANet, we add the multi-scale-aware modules to SFANet architecture for better tackling drastic scale changes of target objects.

- Furthermore, the decoder structure of M-SFANet is adjusted to have more residual connections in order to ensure that the learned multi-scale features of high-level semantic information will impact how the model regress for the final density map.

- For M-SegNet, we change the up-sampling algorithm from bilinear to max unpooling using the memorized indices employed in SegNet. This yields the cheaper computation model while providing competitive counting performance applicable to real-world applications.
Selected references


Thank you for your attention!