Learning Adaptive Receptive Fields for Deep Image Parsing Network
Zhen Wei¹,⁴, Yao Sun¹, Jinqiao Wang², Hanjiang Lai³, Si Liu¹
1. IIE, CAS, Beijing, China 2. IA, CAS, Beijing, China 3. SYU, Guangzhou, China 4. UCAS, Beijing, China

Introduction:
➢ In this paper, we propose a learning based, data-driven method for regulating receptive field in deep image parsing network automatically.

Framework:
• Two derivable affine transformation layers are introduced into the network;
• The new layers are inserted before (inflation layer) and after (interpolation layer) the convolutional layer whose receptive fields need to be regulated (e.g. fc6 – fc8 layers);

Introduction:
➢ In this paper, we propose a learning based, data-driven method for regulating receptive field in deep image parsing network automatically.

Framework:
• Two derivable affine transformation layers are introduced into the network;
• The new layers are inserted before (inflation layer) and after (interpolation layer) the convolutional layer whose receptive fields need to be regulated (e.g. fc6 – fc8 layers);

Method:
➢ Detailed explanation of the two affine transformation layers:
• Forward (affine transformation on feature maps):
  – inflation layer: resize each feature maps with the factor of $f$;
  – interpolation layer: resize feature maps back to a fixed size, the factor $f' = f/F$, where $F$ is a pre-defined constant.
• Backward (the gradient computation of $f$):
  – inflation layer: the gradient w.r.t. $f$: $\frac{\partial L}{\partial f} = \frac{\partial L}{\partial G_{in}} \cdot \frac{\partial G_{in}}{\partial f}$.
  – interpolation layer: the gradient w.r.t. $f$: $\frac{\partial L}{\partial f} = \frac{\partial L}{\partial G_{inter}} \cdot \frac{\partial G_{inter}}{\partial f}$
• In implementation, the two gradients of $f$ are added together:
  \[ \frac{\partial L}{\partial f} = \frac{\partial L}{\partial G_{in}} \cdot \frac{\partial G_{in}}{\partial f} + \frac{\partial L}{\partial G_{inter}} \cdot \frac{\partial G_{inter}}{\partial f} \]

Experiments:
➢ Here we present results from the single-path networks trained on VOC dataset (general image parsing task) and Helen dataset (face parsing task).

Image parsing results:

Table: Image parsing results
<table>
<thead>
<tr>
<th>Image</th>
<th>gt</th>
<th>best manual rf</th>
<th>improper initial rf</th>
<th>improper rf only</th>
</tr>
</thead>
<tbody>
<tr>
<td>image</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>