Learning Salient Objects in a Scene using Superpixel-augmented Convolutional Neural Network
Problem Overview

Given a natural scene image, objective is to detect “salient” regions.
1. **Superpixels are generated using SLIC segmentation algorithm**
2. **Saliency detection specific Hand-crafted representations are obtained**
3. **Per superpixel binary classification is done using shallow CNN**
SLIC superpixel generation

\[ d_c = \sqrt{(l_j - l_i)^2 + (a_j - a_i)^2 + (b_j - b_i)^2} \]

\[ d_s = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} \]

\[ D' = \sqrt{\left(\frac{d_c}{N_c}\right)^2 + \left(\frac{d_s}{N_s}\right)^2} \]
Superpixel representations

\[
Q = \begin{bmatrix}
q_{11}^c & \cdots & q_{1j}^c & \cdots & q_{1M}^c \\
\vdots & & \vdots & & \vdots \\
q_{x1}^c & \cdots & q_{xj}^c & \cdots & q_{xM}^c \\
\vdots & & \vdots & & \vdots \\
q_{N1}^c & \cdots & q_{Nj}^c & \cdots & q_{NM}^c
\end{bmatrix}
\]

\[
Q' = \begin{bmatrix}
q_{11}^d & \cdots & q_{1j}^d & \cdots & q_{1M}^d \\
\vdots & & \vdots & & \vdots \\
q_{x1}^d & \cdots & q_{xj}^d & \cdots & q_{xM}^d \\
\vdots & & \vdots & & \vdots \\
q_{N1}^d & \cdots & q_{Nj}^d & \cdots & q_{NM}^d
\end{bmatrix}
\]

\[q_{xj}^c = t(r_j) \cdot |C(r_x) - C(r_j)| \cdot w(P(r_x), P(r_j))\]

\[q_{xj}^d = t(r_j) \cdot |P(r_x) - P(r_j)| \cdot w(C(r_x), C(r_j))\]
**Classification using shallow CNN**

<table>
<thead>
<tr>
<th>Type</th>
<th>Size/Stride</th>
<th>Number of channels</th>
</tr>
</thead>
<tbody>
<tr>
<td>input</td>
<td>M</td>
<td>3</td>
</tr>
<tr>
<td>Conv1D, tanh</td>
<td>10 × 1</td>
<td>5</td>
</tr>
<tr>
<td>MaxPool</td>
<td>2 × 1/2 × 1</td>
<td>-</td>
</tr>
<tr>
<td>Conv1D, tanh</td>
<td>20 × 1</td>
<td>10</td>
</tr>
<tr>
<td>Maxpool</td>
<td>2 × 1/2 × 1</td>
<td>-</td>
</tr>
<tr>
<td>Conv1D, tanh</td>
<td>20 × 1</td>
<td>20</td>
</tr>
<tr>
<td>Maxpool</td>
<td>2 × 1/2 × 1</td>
<td>-</td>
</tr>
<tr>
<td>Global-avg-pool</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Fully-con, tanh</td>
<td>20</td>
<td>-</td>
</tr>
<tr>
<td>Dropout (0.3)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Fully-con, Softmax</td>
<td>2</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1: CNN architecture used in this project. Note: size of Q matrix is M×M.

**Total number of parameters:** 5,233

**Total Number of parameters in conv layers of vgg-16:** 14,714,688
Results
Conclusions

1. **Hand-crafted features capturing global spatial context** at superpixel level along with CNN classifier shows decent performance for saliency detection in images.

2. **Superpixel-CNN based approach** does not require large training data and number of parameters in CNN are in *few thousands* compared to *10s of millions* of parameters in standard CNNs on raw image pixels.

3. **Lower number of parameters** reduces the computational cost *significantly* at both training and test time.
THANKS!

Shashank Tripathi
Yash Patel

shashanktripathi123@github
yash0307@github

Code is available here: https://github.com/yash0307/SuperCNN