SAM: Pushing the Limits of Saliency Prediction Models
Marcella Cornia¹, Lorenzo Baraldi¹, Giuseppe Serra² and Rita Cucchiara¹
¹University of Modena and Reggio Emilia, Italy ²University of Udine, Italy

Saliency ATTENTION Model (SAM)

Attentive ConvLSTM

• Extension of the traditional LSTM to work on spatial features by substituting dot products with convolutional operations.
• Exploitation of the sequential nature of the LSTM to process features in an iterative way, without the concept of time.

The input of the LSTM layer \( \mathbf{x}_t \) is computed through an attention mechanism which produces an attention map from the previous hidden state \( \mathbf{h}_{t-1} \) of the LSTM and the input \( \mathbf{x} \)

\[
\mathbf{z}_t = \mathbf{v}_a \cdot \text{tanh}(\mathbf{W}_a \cdot \mathbf{x}_t + \mathbf{u}_a \cdot \mathbf{h}_{t-1} + \mathbf{b}_a)
\]

The output of this operation is a 2-d map from which we compute a normalized spatial attention map \( \mathbf{a}_i \) through the softmax operator.

The attention map is applied to the input with an element-wise product between each channel of the feature maps and the attention map.

\[
\mathbf{x}_t = \mathbf{a}_i \odot \mathbf{x}
\]

Attentive ConvLSTM

Dilated Convolutional Network

We build two different versions of our model based on VGG-16 and ResNet-50.

Progressive refinement of saliency maps

\[
\begin{align*}
\text{NSS} & \quad \text{CC} & \quad \text{GT}
\end{align*}
\]

Dilated Convolutional Network

• To take different quality aspects into account, we define a new loss function given by a linear combination of three saliency evaluation metrics: the NSS, the CC and the KL-Div.

\[
\begin{align*}
\text{Loss Function} = \lambda_1 \times \text{NSS} + \lambda_2 \times \text{CC} + \lambda_3 \times \text{KL-Div}
\end{align*}
\]

To limit rescaling, we use dilated convolutions thus obtaining saliency maps rescaled by a factor of 8 instead of 32.

Learned Priors

• Our network is able to learn the center bias present in eye fixations, without integrating this information manually.

• The model learns means and variances of a set of Gaussian functions with diagonal covariance matrix and produces a prior map for each function.

Results on SALICON 2015

<table>
<thead>
<tr>
<th>Model</th>
<th>CC</th>
<th>AUC</th>
<th>NSS</th>
<th>GT</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAM-Bilinear</td>
<td>0.84</td>
<td>0.85</td>
<td>3.20</td>
<td>0.90</td>
</tr>
<tr>
<td>SAM-VGG</td>
<td>0.83</td>
<td>0.85</td>
<td>3.14</td>
<td>0.90</td>
</tr>
<tr>
<td>U-Net</td>
<td>0.73</td>
<td>0.67</td>
<td>2.79</td>
<td>0.87</td>
</tr>
<tr>
<td>SaltAN [2]</td>
<td>0.78</td>
<td>0.78</td>
<td>2.46</td>
<td>0.84</td>
</tr>
<tr>
<td>SaltAN [2]</td>
<td>0.62</td>
<td>0.68</td>
<td>1.86</td>
<td>0.76</td>
</tr>
<tr>
<td>DeepGazeII [1]</td>
<td>0.51</td>
<td>0.60</td>
<td>1.31</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Results on SALICON 2017

<table>
<thead>
<tr>
<th>Model</th>
<th>CC</th>
<th>AUC</th>
<th>NSS</th>
<th>GT</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAM-Bilinear</td>
<td>0.80</td>
<td>0.87</td>
<td>1.09</td>
<td>0.90</td>
</tr>
<tr>
<td>SAM-VGG</td>
<td>0.89</td>
<td>0.86</td>
<td>1.97</td>
<td>0.90</td>
</tr>
<tr>
<td>U-Net</td>
<td>0.77</td>
<td>0.68</td>
<td>2.79</td>
<td>0.87</td>
</tr>
<tr>
<td>SaltAN [2]</td>
<td>0.84</td>
<td>0.86</td>
<td>1.82</td>
<td>0.76</td>
</tr>
<tr>
<td>SaltAN [2]</td>
<td>0.76</td>
<td>0.68</td>
<td>1.46</td>
<td>0.56</td>
</tr>
<tr>
<td>DeepGazeIIII [1]</td>
<td>0.51</td>
<td>0.60</td>
<td>1.31</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Experimental Results

First place in the LSUN Saliency Prediction Challenge (CVPR 2019)

State of the art on both versions of SALICON, the largest dataset available for saliency.

Very good results on several other datasets such as MIT300, MIT1003, CAT2000, DUT-OMRON, TORONTO and PASCAL-S.

References: