Understanding Deep Neural Networks
Via Smooth Masks and Extremal Perturbations

Ruth Fong*, Mandela Patrick*, Andrea Vedaldi
ICCV 2019

* equal contribution
Attribution

Where is the model looking?
Backpropogation

Combine network activations and gradients

Input  Gradient  GRAD-CAM

Fast, but difficult to characterize

[Simonyan et al., ICLR Workshop 2014; Selvaraju et al., ICCV 2017]
[Mahendran and Vedaldi, ECCV 2016; Adebayo et al., NeurIPS 2018]
Perturbation

Change the input and observe the effect on the output

Clear meaning, but can only test a small number of occlusion patterns

[Zeiler and Fergus, ECCV 2014; Petsiuk et al., BMVC 2018]
Blur everywhere => response suppressed
Preserve 10% => response preserved
Extremal Perturbations

A mask is optimized to maximally excite the network:

$$\text{argmax}_m \Phi(m \otimes x)$$

subject to area(m) = a
Area Constraint

Optimizing w.r.t. to an area constraint is challenging

Here we re-formulate it as matching a rank statistics

\[ L_{\text{area}} = \| \text{vecsort}(m) - r_a \|^2 \]

Concurrent work: [Kapishnikov et al., arXiv 2019]
Smooth Masks

\[
\text{conv}(u; m; k) = \frac{1}{Z} \sum_{v \in \Omega} k(u - v)m(v)
\]

\[
\text{maxconv}(u; m; k) = \max_{v \in \Omega} k(u - v)m(v)
\]

\[
\text{smoothconv}(u; m; k; T) = \text{smax}_{v \in \Omega; T} k(u - v)m(v)
\]
Smooth Masks
Algorithm

1. Pick an area $a$

2. Use SGD to solve the optimization problem for a large $\lambda$:

$$\arg\max_m \Phi(\text{smooth}(m) \otimes x) - \lambda \| \text{vecsort}(\text{smooth}(m)) - r_a \|^2$$

3. If needed, sweep $a$ and repeat
Comparison to prior work on “meaningful perturbations”
Results
Foreground evidence is usually sufficient
Large objects are recognised by their details
Multiple objects contribute cumulatively
Pointing game: weak localization

<table>
<thead>
<tr>
<th>Method</th>
<th>( \text{VOC07 Test (All/Diff)} )</th>
<th>( \text{COCO14 Val (All/Diff)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \text{VGG16} )</td>
<td>( \text{ResNet50} )</td>
</tr>
<tr>
<td>Grad</td>
<td>76.3/56.9</td>
<td>72.3/56.8</td>
</tr>
<tr>
<td>DConv</td>
<td>67.5/44.1</td>
<td>68.6/44.7</td>
</tr>
<tr>
<td>Guid.</td>
<td>75.9/53.0</td>
<td>77.2/59.5</td>
</tr>
<tr>
<td>MWP</td>
<td>77.1/56.6</td>
<td>84.4/70.8</td>
</tr>
<tr>
<td>cMWP</td>
<td>79.9/66.5</td>
<td>\textbf{90.6/82.2}</td>
</tr>
<tr>
<td>RISE*</td>
<td>87.3/—</td>
<td>88.9/—</td>
</tr>
<tr>
<td>GCAM</td>
<td>86.6/74.0</td>
<td>\textbf{90.4/82.3}</td>
</tr>
<tr>
<td>Ours</td>
<td>\textbf{88.7/75.5}</td>
<td>86.3/73.4</td>
</tr>
</tbody>
</table>

[Zhang et al., ECCV 2016]
Attributing channels at intermediate layers
Spatial Attribution
Channel Attribution

\[ x \xrightarrow{\text{perturb}} \Phi_a, \Phi_b \xrightarrow{\Phi (m \otimes x)} \]
Channel Attribution

\[
\Phi_{a}(x) \xrightarrow{\text{perturb}} m \otimes \Phi_{a}(x) \xrightarrow{\Phi_{b}} \Phi_{b}(m \otimes \Phi_{a}(x))
\]

argmax \( \Phi_{b}(m \otimes \Phi_{a}(x)) \)

subject to \( \text{area}(m) = a \)
Activation “diffing”

Original $\Phi_a(x)$

Perturbed $m \otimes \Phi_a(x)$

[Olah et al., Distill 2017]