Bias Attribution for Deep Neural Networks.

<table>
<thead>
<tr>
<th>label</th>
<th>original</th>
<th>nom. grad.</th>
<th>grad. attrib.</th>
<th>nom. integrad.</th>
<th>integrad. attrib.</th>
<th>nom. bias.1</th>
<th>bias.1 attrib.</th>
<th>nom. bias.2</th>
<th>bias.2 attrib.</th>
<th>nom. bias.3</th>
<th>bias.3 attrib.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dhole</td>
<td><img src="Dhole.png" alt="Image" /></td>
<td><img src="Dhole_nom_grad.png" alt="Image" /></td>
<td><img src="Dhole_grad_attrib.png" alt="Image" /></td>
<td><img src="Dhole_nom_integra.png" alt="Image" /></td>
<td><img src="Dhole_integrad_attrib.png" alt="Image" /></td>
<td><img src="Dhole_bias_1.png" alt="Image" /></td>
<td><img src="Dhole_bias_1_attrib.png" alt="Image" /></td>
<td><img src="Dhole_bias_2.png" alt="Image" /></td>
<td><img src="Dhole_bias_2_attrib.png" alt="Image" /></td>
<td><img src="Dhole_bias_3.png" alt="Image" /></td>
<td><img src="Dhole_bias_3_attrib.png" alt="Image" /></td>
</tr>
<tr>
<td>Kite (bird)</td>
<td><img src="Kite.png" alt="Image" /></td>
<td><img src="Kite_nom_grad.png" alt="Image" /></td>
<td><img src="Kite_grad_attrib.png" alt="Image" /></td>
<td><img src="Kite_nom_integra.png" alt="Image" /></td>
<td><img src="Kite_integrad_attrib.png" alt="Image" /></td>
<td><img src="Kite_bias_1.png" alt="Image" /></td>
<td><img src="Kite_bias_1_attrib.png" alt="Image" /></td>
<td><img src="Kite_bias_2.png" alt="Image" /></td>
<td><img src="Kite_bias_2_attrib.png" alt="Image" /></td>
<td><img src="Kite_bias_3.png" alt="Image" /></td>
<td><img src="Kite_bias_3_attrib.png" alt="Image" /></td>
</tr>
<tr>
<td>Angora Rabbit</td>
<td><img src="Angora.png" alt="Image" /></td>
<td><img src="Angora_nom_grad.png" alt="Image" /></td>
<td><img src="Angora_grad_attrib.png" alt="Image" /></td>
<td><img src="Angora_nom_integra.png" alt="Image" /></td>
<td><img src="Angora_integrad_attrib.png" alt="Image" /></td>
<td><img src="Angora_bias_1.png" alt="Image" /></td>
<td><img src="Angora_bias_1_attrib.png" alt="Image" /></td>
<td><img src="Angora_bias_2.png" alt="Image" /></td>
<td><img src="Angora_bias_2_attrib.png" alt="Image" /></td>
<td><img src="Angora_bias_3.png" alt="Image" /></td>
<td><img src="Angora_bias_3_attrib.png" alt="Image" /></td>
</tr>
<tr>
<td>Cicada</td>
<td><img src="Cicada.png" alt="Image" /></td>
<td><img src="Cicada_nom_grad.png" alt="Image" /></td>
<td><img src="Cicada_grad_attrib.png" alt="Image" /></td>
<td><img src="Cicada_nom_integra.png" alt="Image" /></td>
<td><img src="Cicada_integrad_attrib.png" alt="Image" /></td>
<td><img src="Cicada_bias_1.png" alt="Image" /></td>
<td><img src="Cicada_bias_1_attrib.png" alt="Image" /></td>
<td><img src="Cicada_bias_2.png" alt="Image" /></td>
<td><img src="Cicada_bias_2_attrib.png" alt="Image" /></td>
<td><img src="Cicada_bias_3.png" alt="Image" /></td>
<td><img src="Cicada_bias_3_attrib.png" alt="Image" /></td>
</tr>
<tr>
<td>Blitern</td>
<td><img src="Blitern.png" alt="Image" /></td>
<td><img src="Blitern_nom_grad.png" alt="Image" /></td>
<td><img src="Blitern_grad_attrib.png" alt="Image" /></td>
<td><img src="Blitern_nom_integra.png" alt="Image" /></td>
<td><img src="Blitern_integrad_attrib.png" alt="Image" /></td>
<td><img src="Blitern_bias_1.png" alt="Image" /></td>
<td><img src="Blitern_bias_1_attrib.png" alt="Image" /></td>
<td><img src="Blitern_bias_2.png" alt="Image" /></td>
<td><img src="Blitern_bias_2_attrib.png" alt="Image" /></td>
<td><img src="Blitern_bias_3.png" alt="Image" /></td>
<td><img src="Blitern_bias_3_attrib.png" alt="Image" /></td>
</tr>
<tr>
<td>Albatross</td>
<td><img src="Albatross.png" alt="Image" /></td>
<td><img src="Albatross_nom_grad.png" alt="Image" /></td>
<td><img src="Albatross_grad_attrib.png" alt="Image" /></td>
<td><img src="Albatross_nom_integra.png" alt="Image" /></td>
<td><img src="Albatross_integrad_attrib.png" alt="Image" /></td>
<td><img src="Albatross_bias_1.png" alt="Image" /></td>
<td><img src="Albatross_bias_1_attrib.png" alt="Image" /></td>
<td><img src="Albatross_bias_2.png" alt="Image" /></td>
<td><img src="Albatross_bias_2_attrib.png" alt="Image" /></td>
<td><img src="Albatross_bias_3.png" alt="Image" /></td>
<td><img src="Albatross_bias_3_attrib.png" alt="Image" /></td>
</tr>
<tr>
<td>King Charles Spaniel</td>
<td><img src="King_Charles_Spaniel.png" alt="Image" /></td>
<td><img src="King_Charles_Spaniel_nom_grad.png" alt="Image" /></td>
<td><img src="King_Charles_Spaniel_grad_attrib.png" alt="Image" /></td>
<td><img src="King_Charles_Spaniel_nom_integra.png" alt="Image" /></td>
<td><img src="King_Charles_Spaniel_integrad_attrib.png" alt="Image" /></td>
<td><img src="King_Charles_Spaniel_bias_1.png" alt="Image" /></td>
<td><img src="King_Charles_Spaniel_bias_1_attrib.png" alt="Image" /></td>
<td><img src="King_Charles_Spaniel_bias_2.png" alt="Image" /></td>
<td><img src="King_Charles_Spaniel_bias_2_attrib.png" alt="Image" /></td>
<td><img src="King_Charles_Spaniel_bias_3.png" alt="Image" /></td>
<td><img src="King_Charles_Spaniel_bias_3_attrib.png" alt="Image" /></td>
</tr>
<tr>
<td>Racket</td>
<td><img src="Racket.png" alt="Image" /></td>
<td><img src="Racket_nom_grad.png" alt="Image" /></td>
<td><img src="Racket_grad_attrib.png" alt="Image" /></td>
<td><img src="Racket_nom_integra.png" alt="Image" /></td>
<td><img src="Racket_integrad_attrib.png" alt="Image" /></td>
<td><img src="Racket_bias_1.png" alt="Image" /></td>
<td><img src="Racket_bias_1_attrib.png" alt="Image" /></td>
<td><img src="Racket_bias_2.png" alt="Image" /></td>
<td><img src="Racket_bias_2_attrib.png" alt="Image" /></td>
<td><img src="Racket_bias_3.png" alt="Image" /></td>
<td><img src="Racket_bias_3_attrib.png" alt="Image" /></td>
</tr>
<tr>
<td>African Wild Dog</td>
<td><img src="African_Wild_Dog.png" alt="Image" /></td>
<td><img src="African_Wild_Dog_nom_grad.png" alt="Image" /></td>
<td><img src="African_Wild_Dog_grad_attrib.png" alt="Image" /></td>
<td><img src="African_Wild_Dog_nom_integra.png" alt="Image" /></td>
<td><img src="African_Wild_Dog_integrad_attrib.png" alt="Image" /></td>
<td><img src="African_Wild_Dog_bias_1.png" alt="Image" /></td>
<td><img src="African_Wild_Dog_bias_1_attrib.png" alt="Image" /></td>
<td><img src="African_Wild_Dog_bias_2.png" alt="Image" /></td>
<td><img src="African_Wild_Dog_bias_2_attrib.png" alt="Image" /></td>
<td><img src="African_Wild_Dog_bias_3.png" alt="Image" /></td>
<td><img src="African_Wild_Dog_bias_3_attrib.png" alt="Image" /></td>
</tr>
<tr>
<td>Standard Poodle</td>
<td><img src="Standard_Poodle.png" alt="Image" /></td>
<td><img src="Standard_Poodle_nom_grad.png" alt="Image" /></td>
<td><img src="Standard_Poodle_grad_attrib.png" alt="Image" /></td>
<td><img src="Standard_Poodle_nom_integra.png" alt="Image" /></td>
<td><img src="Standard_Poodle_integrad_attrib.png" alt="Image" /></td>
<td><img src="Standard_Poodle_bias_1.png" alt="Image" /></td>
<td><img src="Standard_Poodle_bias_1_attrib.png" alt="Image" /></td>
<td><img src="Standard_Poodle_bias_2.png" alt="Image" /></td>
<td><img src="Standard_Poodle_bias_2_attrib.png" alt="Image" /></td>
<td><img src="Standard_Poodle_bias_3.png" alt="Image" /></td>
<td><img src="Standard_Poodle_bias_3_attrib.png" alt="Image" /></td>
</tr>
<tr>
<td>Waffle Iron</td>
<td><img src="Waffle_Iron.png" alt="Image" /></td>
<td><img src="Waffle_Iron_nom_grad.png" alt="Image" /></td>
<td><img src="Waffle_Iron_grad_attrib.png" alt="Image" /></td>
<td><img src="Waffle_Iron_nom_integra.png" alt="Image" /></td>
<td><img src="Waffle_Iron_integrad_attrib.png" alt="Image" /></td>
<td><img src="Waffle_Iron_bias_1.png" alt="Image" /></td>
<td><img src="Waffle_Iron_bias_1_attrib.png" alt="Image" /></td>
<td><img src="Waffle_Iron_bias_2.png" alt="Image" /></td>
<td><img src="Waffle_Iron_bias_2_attrib.png" alt="Image" /></td>
<td><img src="Waffle_Iron_bias_3.png" alt="Image" /></td>
<td><img src="Waffle_Iron_bias_3_attrib.png" alt="Image" /></td>
</tr>
<tr>
<td>Starfish</td>
<td><img src="Starfish.png" alt="Image" /></td>
<td><img src="Starfish_nom_grad.png" alt="Image" /></td>
<td><img src="Starfish_grad_attrib.png" alt="Image" /></td>
<td><img src="Starfish_nom_integra.png" alt="Image" /></td>
<td><img src="Starfish_integrad_attrib.png" alt="Image" /></td>
<td><img src="Starfish_bias_1.png" alt="Image" /></td>
<td><img src="Starfish_bias_1_attrib.png" alt="Image" /></td>
<td><img src="Starfish_bias_2.png" alt="Image" /></td>
<td><img src="Starfish_bias_2_attrib.png" alt="Image" /></td>
<td><img src="Starfish_bias_3.png" alt="Image" /></td>
<td><img src="Starfish_bias_3_attrib.png" alt="Image" /></td>
</tr>
<tr>
<td>Hot Dog</td>
<td><img src="Hot_Dog.png" alt="Image" /></td>
<td><img src="Hot_Dog_nom_grad.png" alt="Image" /></td>
<td><img src="Hot_Dog_grad_attrib.png" alt="Image" /></td>
<td><img src="Hot_Dog_nom_integra.png" alt="Image" /></td>
<td><img src="Hot_Dog_integrad_attrib.png" alt="Image" /></td>
<td><img src="Hot_Dog_bias_1.png" alt="Image" /></td>
<td><img src="Hot_Dog_bias_1_attrib.png" alt="Image" /></td>
<td><img src="Hot_Dog_bias_2.png" alt="Image" /></td>
<td><img src="Hot_Dog_bias_2_attrib.png" alt="Image" /></td>
<td><img src="Hot_Dog_bias_3.png" alt="Image" /></td>
<td><img src="Hot_Dog_bias_3_attrib.png" alt="Image" /></td>
</tr>
<tr>
<td>Hen</td>
<td><img src="Hen.png" alt="Image" /></td>
<td><img src="Hen_nom_grad.png" alt="Image" /></td>
<td><img src="Hen_grad_attrib.png" alt="Image" /></td>
<td><img src="Hen_nom_integra.png" alt="Image" /></td>
<td><img src="Hen_integrad_attrib.png" alt="Image" /></td>
<td><img src="Hen_bias_1.png" alt="Image" /></td>
<td><img src="Hen_bias_1_attrib.png" alt="Image" /></td>
<td><img src="Hen_bias_2.png" alt="Image" /></td>
<td><img src="Hen_bias_2_attrib.png" alt="Image" /></td>
<td><img src="Hen_bias_3.png" alt="Image" /></td>
<td><img src="Hen_bias_3_attrib.png" alt="Image" /></td>
</tr>
<tr>
<td>Trifle</td>
<td><img src="Trifle.png" alt="Image" /></td>
<td><img src="Trifle_nom_grad.png" alt="Image" /></td>
<td><img src="Trifle_grad_attrib.png" alt="Image" /></td>
<td><img src="Trifle_nom_integra.png" alt="Image" /></td>
<td><img src="Trifle_integrad_attrib.png" alt="Image" /></td>
<td><img src="Trifle_bias_1.png" alt="Image" /></td>
<td><img src="Trifle_bias_1_attrib.png" alt="Image" /></td>
<td><img src="Trifle_bias_2.png" alt="Image" /></td>
<td><img src="Trifle_bias_2_attrib.png" alt="Image" /></td>
<td><img src="Trifle_bias_3.png" alt="Image" /></td>
<td><img src="Trifle_bias_3_attrib.png" alt="Image" /></td>
</tr>
<tr>
<td>Scottish Deerhound</td>
<td><img src="Scottish_Deerhound.png" alt="Image" /></td>
<td><img src="Scottish_Deerhound_nom_grad.png" alt="Image" /></td>
<td><img src="Scottish_Deerhound_grad_attrib.png" alt="Image" /></td>
<td><img src="Scottish_Deerhound_nom_integra.png" alt="Image" /></td>
<td><img src="Scottish_Deerhound_integrad_attrib.png" alt="Image" /></td>
<td><img src="Scottish_Deerhound_bias_1.png" alt="Image" /></td>
<td><img src="Scottish_Deerhound_bias_1_attrib.png" alt="Image" /></td>
<td><img src="Scottish_Deerhound_bias_2.png" alt="Image" /></td>
<td><img src="Scottish_Deerhound_bias_2_attrib.png" alt="Image" /></td>
<td><img src="Scottish_Deerhound_bias_3.png" alt="Image" /></td>
<td><img src="Scottish_Deerhound_bias_3_attrib.png" alt="Image" /></td>
</tr>
</tbody>
</table>

Figure 5: More visualizations of bias attribution on the ImageNet compared to gradient and integrated gradient attribution.
<table>
<thead>
<tr>
<th>Label</th>
<th>Original</th>
<th>norm. grad.</th>
<th>grad. attrib.</th>
<th>norm. integrat.</th>
<th>attrib.</th>
<th>norm. bias.1</th>
<th>bias.1</th>
<th>norm. bias.2</th>
<th>bias.2</th>
<th>norm. bias.3</th>
<th>bias.3</th>
<th>integrad.</th>
<th>integrad.</th>
<th>attrib.</th>
</tr>
</thead>
</table>

Figure 6: More visualizations of bias attribution on the ImageNet compared to gradient and integrated gradient attribution.
A MNIST Digit Flip Test Network Details

The network we use is a slight modification of the lenet structure (see Table 2). The trained network achieves 91.1% accuracy on the test set. For the digit flip test, we remove 120 pixels from each image to convert from the source class to the target class based on the descending order of source class attribution scores minus target class attribution scores. For the bias attribution methods, since the attribution scores add up to the bias term, which is different for different classes, we normalize the bias attribution scores by their L2 norm.

<table>
<thead>
<tr>
<th>Layer Type</th>
<th>Layer Size (kernel size, stride, padding, channels) / # hidden units</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv1</td>
<td>[3 × 3], 1, 1, 32</td>
</tr>
<tr>
<td>Batchnorm</td>
<td>N/A</td>
</tr>
<tr>
<td>ReLU</td>
<td>N/A</td>
</tr>
<tr>
<td>conv2</td>
<td>[3 × 3], 1, 1, 32</td>
</tr>
<tr>
<td>Batchnorm</td>
<td>N/A</td>
</tr>
<tr>
<td>ReLU</td>
<td>N/A</td>
</tr>
<tr>
<td>MaxPool</td>
<td>[2 × 2], 2, 0, N/A</td>
</tr>
<tr>
<td>Fully Connected</td>
<td>128</td>
</tr>
<tr>
<td>Batchnorm</td>
<td>N/A</td>
</tr>
<tr>
<td>ReLU</td>
<td>N/A</td>
</tr>
<tr>
<td>Fully Connected</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 2: Neural network