InsideBias: Measuring Bias in Deep Networks and Application to Face Gender Biometrics

MSc. Ignacio SERNA
MSc. Alejandro PEÑA
Dr. Aythami MORALES
http://www.aythami.me
Prof. Julian FIERREZ
http://biometrics.eps.uam.es
Biases in Deep Networks

AI trend during last decade: **excellent performance** + low transparency

**Biased Training Process**

**Biased Performance**
Contributions

• Comprehensive analysis of biased learning demonstrating the latent **correlations** between **bias**, **activation** of Neural Networks, and recognition **performance**.

  ![Bias Activation Performance Diagram]

  Bias ➔ Activations ➔ Performance

• We propose **InsideBias**, a novel **bias detection** method based on the analysis of the filter's activation of deep networks.
Our Method: InsideBias

Convolutional Neural Network (CNN)

- **Input**
- **Filters**
- **1st layer**
- **2nd layer**
- **3rd layer**
- **4th layer**
- **...th layer**
- **Last layer**

- **Convolution**
- **Feature extraction**

- **Fully connected**
- **Classification**

**Examples:**
- Airplane
- Cat
- Bike
- Dog

BiDA Lab
Our Method: InsideBias

- Filters at layer $l = 1$
Our Method: InsideBias

- **Step 1.** The input is *convolved* with the *filters* of the layer $l$ and put through the activation function to form the output *feature maps* $A_i^{[l]}$. 
Our Method: InsideBias

**Step 2.** We compute the average activation $A_i^{[l]}$ for each feature map $i$ in the layer $l$:

$$\text{Average Activation} = \overline{A_i^{[l]}} = 2.10$$
Our Method: InsideBias

**Step 2.** We compute the average activation $A_i^{[l]}$ for each feature map $i$ in the layer $l$:

$$\text{Average Activation} = \frac{A_i^{[l]}}{} = 2.10$$

<table>
<thead>
<tr>
<th></th>
<th>0.11</th>
<th>1.34</th>
<th>2.02</th>
<th>3.17</th>
<th>0.33</th>
<th>6.54</th>
<th>0.56</th>
<th>3.78</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.51</td>
<td>1.56</td>
<td>2.37</td>
<td>2.89</td>
<td>2.21</td>
<td>0.10</td>
<td>0.80</td>
<td>2.51</td>
<td></td>
</tr>
<tr>
<td>2.68</td>
<td>3.24</td>
<td>3.35</td>
<td>0.42</td>
<td>2.26</td>
<td>5.84</td>
<td>1.78</td>
<td>1.40</td>
<td></td>
</tr>
<tr>
<td>5.99</td>
<td>4.04</td>
<td>3.89</td>
<td>3.23</td>
<td>4.95</td>
<td>4.02</td>
<td>1.98</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>0.11</td>
<td>1.78</td>
<td>3.25</td>
<td>3.52</td>
<td>4.73</td>
<td>1.94</td>
<td>3.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.30</td>
<td>0.84</td>
<td>3.79</td>
<td>2.81</td>
<td>5.71</td>
<td>5.86</td>
<td>2.98</td>
<td>3.04</td>
<td></td>
</tr>
<tr>
<td>6.08</td>
<td>5.76</td>
<td>0.08</td>
<td>1.71</td>
<td>0.92</td>
<td>1.23</td>
<td>0.12</td>
<td>1.54</td>
<td></td>
</tr>
<tr>
<td>1.67</td>
<td>4.55</td>
<td>0.09</td>
<td>4.81</td>
<td>4.67</td>
<td>4.44</td>
<td>4.53</td>
<td>1.76</td>
<td></td>
</tr>
</tbody>
</table>
**Our Method: InsideBias**

- **Step 3.** We calculate the *activation* $\lambda^{[l]}$ as the *maximum* of $A_i^{[l]}$ for all feature maps in layer $l$.

\[
\begin{bmatrix}
0.11 & 1.34 & 2.02 & 3.17 & 0.33 & 0.56 & 3.78 \\
6.51 & 1.56 & 2.37 & 2.89 & 2.21 & 0.10 & 0.80 & 2.51 \\
2.68 & 3.24 & 3.35 & 0.42 & 2.26 & 5.84 & 1.78 & 1.40 \\
5.99 & 4.04 & 3.89 & 3.23 & 4.95 & 4.02 & 1.98 & 0.03 \\
0.11 & 1.78 & 2.10 & 3.25 & 3.52 & 4.73 & 1.94 & 3.17 \\
6.30 & 0.84 & 3.79 & 2.81 & 5.71 & 5.86 & 2.98 & 3.04 \\
6.08 & 5.76 & 0.08 & 1.71 & 0.92 & 1.23 & 0.12 & 1.54 \\
1.67 & 4.55 & 0.09 & 4.81 & 4.67 & 4.44 & 4.53 & 1.76 \\
\end{bmatrix}
\]

\[
\lambda^{[l]} = \max_i \left( A_i^{[l]} \right) = 6.54
\]
Our Method: InsideBias

Convolutional Neural Network (CNN)
Experimental Framework

- **Two Tasks:**
  - **Digit recognition**
    - **Architecture:** \textbf{VGG} (610k parameters)
    - **Dataset:** \textbf{Colored MNIST} (inspired by [1])
      - 3 RGB colors (red, green and blue).
      - Training set: 60k samples. 90% primary color and 10% remaining two colors.
      - Test set: 10k samples. Equal color distribution (1/3 x 3).
  - **Face gender classification**
    - **Architecture:** \textbf{VGG} (660k) and \textbf{ResNet} (370k)
    - **Dataset:** \textbf{DiveFace} (available in [2])
      - Gender
        - Male
        - Female
      - Ethnicity
        - A. Japan, China, Korea, ...
        - B. Sub-Saharan Africa, India, ...
        - C. Europe, America, ...

---

Experiments: Digit Recognition

Average $\lambda^{[l]}$ activation in the last convolutional layer for the digits of each color

Green Biased Model

![Green Biased Model Graph]

Unbiased Model

![Unbiased Model Graph]
Experiments: Face Gender Recognition

- **Biased**: priviledged ethnic group = 90% | other two = 5%
- **Unbiased**: 33% images from each group

<table>
<thead>
<tr>
<th>Model</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Avg</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet Biased (A)</td>
<td>96.84</td>
<td>94.14</td>
<td>94.45</td>
<td>95.14</td>
<td>1.21</td>
</tr>
<tr>
<td>ResNet Biased (B)</td>
<td>93.29</td>
<td>96.86</td>
<td>95.40</td>
<td>95.18</td>
<td>1.47</td>
</tr>
<tr>
<td>ResNet Biased (C)</td>
<td>94.80</td>
<td>95.21</td>
<td>97.01</td>
<td>95.67</td>
<td>0.96</td>
</tr>
<tr>
<td>ResNet Unbiased</td>
<td>95.50</td>
<td>95.35</td>
<td>96.11</td>
<td>95.65</td>
<td>0.33</td>
</tr>
</tbody>
</table>

* Database available: [https://github.com/BiDAlab/DiveFace](https://github.com/BiDAlab/DiveFace)

Experiments: Face Gender Recognition

Model: ResNet Biased (A)

Model: ResNet Biased (B)

Model: ResNet Biased (C)

Normalized activation $\lambda'$ observed for the three demographic Groups (A, B and C) in the different trained models.
Experiments: Detecting Bias with Very Few Samples

Only 5 samples

Group A

Biased Model A

Confidence score (in Gender Classification)

Activation $\lambda^{[l]}$ (for the Group)

Activation Ratio $\Lambda^{[l]}$

100% 2.82

Group B

Biased Model A

100% 2.65

2.53 / 2.82 = 0.90

Group C

Biased Model A

100% 2.53
Conclusions

• We have proposed a method for bias detection in deep networks through filter activation.
• The method has been tested in two databases: colored MNIST and DiveFace for handwritten digit and face gender classification.
• The results show that activations are good indicators of the bias introduced during training.

KNOW MORE:


