White-box Fairness Testing through Adversarial Sampling

Peixin Zhang\textsuperscript{1,3}, Jingyi Wang\textsuperscript{1,2*}, Jun Sun\textsuperscript{3},
Guoliang Dong\textsuperscript{1,3}, Xinyu Wang\textsuperscript{1*}, Xingen Wang\textsuperscript{1}, Jinsong Dong\textsuperscript{2}, Ting Dai\textsuperscript{4}

\textsuperscript{1}Zhejiang University
\textsuperscript{2}National University of Singapore
\textsuperscript{3}Singapore Management University
\textsuperscript{4}Huawei International Pte Ltd

2020.07.07
Why Fairness

I CAN'T BREATHE
Individual Discrimination

Given \( x = \{x_1, x_2, \ldots, x_n\} \) where \( x_i \) is the value of attribute \( A_i \) in its domain \( I_i \), and protected attributes \( P \subseteq A \). Say that \( x \) is an individual discriminatory instance (IDI) of a model \( D \) if:

- \( \exists p \in P, \text{ s.t., } x_p \neq x'_p \)
- \( \forall q \in NP, x_q = x'_q \)
- \( D(x) \neq D(x') \)

**Testing:** how can we effectively and efficiently generate IDIs for a given model with potential bias?

Example: “Being male is vile.” versus “Being female is vile.”
Existing Heuristics

• THEMIS (FSE'17)
  • Random without any guide.

• AEQUITAS (ASE'18)
  • Two of three local methods are guided.
  • Guide is not input specific.

• Symbolic Generation (FSE'19)
  • Combine model explanation and symbolic execution.
  • Heavyweight.

Can we propose a better algorithm specifically for deep learning models?
Intuition

Adversarial Attack.

Fairness Testing.
Adversarial Discrimination Finder (ADF)
Global

Problem 1: How to improve the diversity of the testing data?
Through clustering.

Problem 2: How do we perturb the data?
Based on the sign of gradients.

Problem 3: How do we filter out the unreal data?
Clip each attribute within its domain.
Local

Problem 1: How do we choose the attribute for local perturbation?
Based on the absolute value of gradients.

Problem 2: How do we filter out the unreal data?
Clip each attribute with its domain.
A Qualitative Comparison

• Our algorithm is **guided by gradient**, which accelerates the discovery of more individual discriminatory instances.

• Our algorithm is **input specific**, which improves the diversity of IDIs.

• Our algorithm is **lightweight**, which makes it more scalable.

<table>
<thead>
<tr>
<th>Feature</th>
<th>THEMIS</th>
<th>AEQUITAS</th>
<th>SG</th>
<th>ADF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guided</td>
<td>✗</td>
<td>✓ (semi)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Input specific</td>
<td>N.A.</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Lightweight</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Evaluation

• Benchmark (tabular)
  • Census Income: age, race, gender
  • German Credit: age, gender
  • Bank Marketing: bank

• Model
  • Six-layer Fully-connected NN

• Research Questions
  • RQ1: How effective is ADF in finding individual discriminatory instance?
  • RQ2: How efficient is ADF in finding individual discriminatory instances?
  • RQ3: How useful are the identified individual discriminatory instances for improving the fairness?
Evaluation

Number of IDIs generated by global generation.

Number of IDIs generated by local generation.

Answer to RQ1: Our algorithm ADF is more effective than state-of-the-art methods.
Evaluation

Time taken to generate 1000 individual discriminatory instances.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Protected Attr.</th>
<th>AEQUITAS</th>
<th>SG</th>
<th>ADF</th>
</tr>
</thead>
<tbody>
<tr>
<td>census</td>
<td>age</td>
<td>172.64</td>
<td>720.49</td>
<td>59.15</td>
</tr>
<tr>
<td>census</td>
<td>race</td>
<td>128.75</td>
<td>506.33</td>
<td>65.95</td>
</tr>
<tr>
<td>census</td>
<td>gender</td>
<td>158.37</td>
<td>2128.42</td>
<td>78.68</td>
</tr>
<tr>
<td>bank</td>
<td>age</td>
<td>191.16</td>
<td>521.79</td>
<td>106.93</td>
</tr>
<tr>
<td>credit</td>
<td>age</td>
<td>176.31</td>
<td>321.63</td>
<td>64.92</td>
</tr>
<tr>
<td>credit</td>
<td>gender</td>
<td>156.22</td>
<td>476.52</td>
<td>102.90</td>
</tr>
</tbody>
</table>

Answer to RQ2: Our algorithm ADF is more efficient than state-of-the-art methods.
Evaluation

Fairness improvement.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Prot. Attr.</th>
<th>Before (%)</th>
<th>After (%)</th>
<th>ADF</th>
<th>AEQUITAS</th>
<th>SG</th>
</tr>
</thead>
<tbody>
<tr>
<td>census</td>
<td>age</td>
<td>10.88</td>
<td></td>
<td>2.26</td>
<td>4.03</td>
<td>2.41</td>
</tr>
<tr>
<td>census</td>
<td>race</td>
<td>9.75</td>
<td></td>
<td>6.15</td>
<td>7.05</td>
<td>6.89</td>
</tr>
<tr>
<td>census</td>
<td>gender</td>
<td>3.14</td>
<td></td>
<td>1.65</td>
<td>2.33</td>
<td>1.90</td>
</tr>
<tr>
<td>bank</td>
<td>age</td>
<td>4.60</td>
<td></td>
<td>1.19</td>
<td>1.68</td>
<td>2.04</td>
</tr>
<tr>
<td>credit</td>
<td>age</td>
<td>27.93</td>
<td></td>
<td>12.05</td>
<td>13.91</td>
<td>13.19</td>
</tr>
<tr>
<td>credit</td>
<td>gender</td>
<td>7.68</td>
<td></td>
<td>3.93</td>
<td>4.58</td>
<td>4.66</td>
</tr>
</tbody>
</table>

Answer to RQ3: The IDIs generated by ADF are useful to improve the fairness of the DNN through retraining.
Conclusion

• We propose a lightweight algorithm to effectively and efficiently generate individual discriminatory instances for deep neural network through adversarial sampling.

• ADF will be expanded beyond structured (tabular) data, e.g., text, image.
Thanks and questions?