Motivation: Deep learning (DL) training is non-deterministic even with a fixed random seed. 2.5% bias difference is caused by DL software implementations (e.g., TensorFlow).

• Floating-point calculations are not associative.
• DL software selects primitive operations at runtime.

Variance affects the evaluation of debiasing algorithms: debiasing algorithms increase bias amplification.

Motivation: Over 80%[1] of DL researchers and practitioners are unaware or unsure about variance in DL model training!

Q26[1] - Do you expect fixed-seed identical DL training runs to be deterministic?

A variance analysis on DL models’ fairness is needed!

Approach: Fairness Variance Analysis

Fixed-Seed Identical Training (FIT) Runs
Using the same
• Random seed
• Training data
• Train test split
• Hyperparameters

Baseline models

Fairness evaluation

Baseline fairness

Statistical analysis (baseline VS debiased)
• Mann-Whitney U-test for mean value
• Levene’s test for variance

Finding: Software alone causes large fairness gap (up to 12.6%)!

<table>
<thead>
<tr>
<th>Technique</th>
<th>Metric</th>
<th>MaxDiff (%)</th>
<th>Max (%)</th>
<th>Min (%)</th>
<th>Avg (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-L2</td>
<td>DP</td>
<td>12.6</td>
<td>39.9</td>
<td>27.3</td>
<td>35.2</td>
</tr>
<tr>
<td>S-GR</td>
<td>DT</td>
<td>11.8</td>
<td>31.1</td>
<td>21.3</td>
<td>28.5</td>
</tr>
</tbody>
</table>

Finding: Hidden cost of debiasing includes higher fairness variance.

About one third (53/154) of the bias mitigation experiments increases fairness variance compared with the corresponding baseline.

Finding: Most (15 out of 22) debiasing techniques increase at least one bias metric.

We call for awareness of implementation-level non-determinism: using proper statistical tests to ensure the validity of deep learning experiments and more!

Reference


Check out our paper here!