Men Also Like Shopping:
Reducing Gender Bias Amplification using Corpus-level Constraints

Jieyu Zhao, Tianlu Wang - University of Virginia
Mark Yatskar - University of Washington
Vincente Ordonez, Kai-Wei Chang - University of Virginia
Motivations

• For vSRL (visual semantic role labeling) and MLC (multi-label object classification)

• Data sets commonly used for above tasks generally contain significant gender bias

• Models trained on these data sets will amplify the gender bias on its predictions

• Desire to reduce the gender bias amplification without reducing accuracy performance.
Previous work:

- Biases in data sets explored
  - Images: (Misra et al. 2016; van Miltenburg, 2016)
  - Text corpus: (Gordon and Van Durme 2013; Van Durme 2010)
- Binary classification
  - (Barocas and Selbst 2014; Dwork et al. 2012)
Bias

Figure 1: Five example images from the imSitu visual semantic role labeling (vSRL) dataset. Each image is paired with a table describing a situation: the verb, cooking, its semantic roles, i.e. agent, and noun values filling that role, i.e. woman. In the imSitu training set, 33% of cooking images have man in the agent role while the rest have woman. After training a Conditional Random Field (CRF), bias is amplified: man fills 16% of agent roles in cooking images. To reduce this bias amplification our calibration method adjusts weights of CRF potentials associated with biased predictions. After applying our methods, man appears in the agent role of 20% of cooking images, reducing the bias amplification by 25%, while keeping the CRF vSRL performance unchanged.
More Bias

agent: woman

agent: man

loading

<table>
<thead>
<tr>
<th>agent</th>
<th>destination</th>
<th>item</th>
<th>tool</th>
<th>place</th>
</tr>
</thead>
<tbody>
<tr>
<td>woman</td>
<td>dishwasher</td>
<td>dish</td>
<td>hand</td>
<td>kitchen</td>
</tr>
</tbody>
</table>

loading

<table>
<thead>
<tr>
<th>agent</th>
<th>destination</th>
<th>item</th>
<th>tool</th>
<th>place</th>
</tr>
</thead>
<tbody>
<tr>
<td>man</td>
<td>truck</td>
<td>bag</td>
<td>forklift</td>
<td>outside</td>
</tr>
</tbody>
</table>

cooking

<table>
<thead>
<tr>
<th>agent</th>
<th>food</th>
<th>container</th>
<th>heatsource</th>
<th>tool</th>
<th>place</th>
</tr>
</thead>
<tbody>
<tr>
<td>woman</td>
<td>vegetable</td>
<td>pan</td>
<td>oven</td>
<td>wooden spoon</td>
<td>kitchen</td>
</tr>
</tbody>
</table>

cooking

<table>
<thead>
<tr>
<th>agent</th>
<th>food</th>
<th>container</th>
<th>heatsource</th>
<th>tool</th>
<th>place</th>
</tr>
</thead>
<tbody>
<tr>
<td>man</td>
<td>Ø</td>
<td>pot</td>
<td>Ø</td>
<td>Ø</td>
<td>Ø</td>
</tr>
</tbody>
</table>
Identifying Bias

Assuming a subset of output variables:

\[ g \subset y, \; g \in G \text{ that reflects demographic attributes} \]

\[ \text{example: } g \in G = \{\text{man, woman, person}\} \]

And another subset of output variables:

\[ o \subset y, \; o \in O \text{ that reflects demographic attributes} \]

where \( o \) is the activity present in the image such as cooking

\[ b(o, g) = \frac{c(o, g)}{\sum_{g' \in G} c(o, g')} \]

\[ \frac{c(verb, man)}{c(verb, man) + c(verb, woman)} \]

If \( b(o, g) > 1/||G|| \), then \( o \) is positively correlated with \( g \) and may exhibit bias.
Measuring bias amplification

\[ \frac{1}{|O|} \sum_{g} \sum_{o \in \{o \in O | b^*(o, g) > 1/\|G'\|\}} \tilde{b}(o, g) - b^*(o, g). \]

\( b^*(o, g) \) = Bias over the training set

\( \tilde{b}(o, g) \) = Bias over the unlabeled evaluation set

This score estimates the average magnitude of bias amplification for pairs of o and g
Calibration algorithm:
**Reducing Bias Amplification**

In essence: Inject constraints to ensure that the model predictions follow the distribution observed from the training data.

Constraints added to vSRL ensure the gender ratio of each verb are within a given margin based on the training dataset. These constraints are added at the corpus level. A joint inference over the test instances is required.

Solving a large inference problem with constraints is hard! Use an approximate inference algorithm based on Lagrangian relaxation.
Structured output prediction

The inference problem: 

$$\arg \max_{y \in Y} f_\theta(y, i),$$

Where $f_\theta(y, i)$ is a scoring function such that:

$$f_\theta(y, i) = \sum_v y_v s_\theta(v, i) + \sum_{v, r} y_{v, r} s_\theta(v, r, i), \quad \text{= overall score of the assignment}$$

In vSRL the output $y$ consists of two binary output variables.

$\{y_{v, r}\}$: = 1 iff activity $v$ and semantic role $r$ are assigned

$\{y_v\}$ = 1 iff activity $v$ is chosen

$s_\theta(v, i)$, $s_\theta(v, r, i)$: Are the potentials of the sub assignments
Corpus-level Constraints

The goal is to inject constraints to ensure the output labels follow a desired distribution

Example: constraint for gender ratio for each activity is within a given margin

$$b^* - \gamma \leq \frac{\sum_i y^i_{v=v^*, r \in M}}{\sum_i y^i_{v=v^*, r \in W} + \sum_i y^i_{v=v^*, r \in M}} \leq b^* + \gamma$$

$$y^i = \{y^i_v\} \cup \{y^i_{v,r}\} = \text{output assignment for test instance } i$$

$$b^* \equiv b^*(v^*, \text{man}) = \text{desired gender ratio for } v^*$$

$$\gamma = \text{user specified margin}$$

In general these corpus level constraints can be represented:

$$A \sum_i y^i - b \leq 0$$
Putting it together

We can reformulate the constrained inference problem as:

\[
\max_{\{y^i\} \in \{Y^i\}} \sum_i f_\theta(y^i, i),
\]

s.t. \[A \sum_i y^i - b \leq 0,\]

Solving a constrained optimization problem on such a large scale is difficult.
This technique can relax the constraints on our original problem.

Introduce a Lagrangian multiplier $\lambda_j \geq 0$ for each corpus-level constraint.

The Lagrangian becomes:

$$L(\lambda, \{y^i\}) = \sum_i f_\theta(y^i) - \sum_{j=1}^l \lambda_j \left( A_j \sum_i y^i - b_j \right),$$

Where: $\lambda_j \geq 0, \forall j \in \{1, \ldots, l\}$
Iterative algorithm

The problem can now be solved by:

1) At iteration $t$, get the output solution of each instance $i$

$$y^{i,(t)} = \arg\max_{y \in \mathcal{Y}} L(\lambda^{(t-1)}, y) \quad (5)$$

2) update the Lagrangian multipliers.

$$\lambda^{(t)} = \max \left( 0, \lambda^{(t-1)} + \sum_i \eta(Ay^{i,(t)} - b) \right),$$

Where:

$$\lambda^{(0)} = 0$$

$\eta$ is the learning rate

This algorithm loops until all constraints are satisfied (an optimal solution is found), or the max number of iterations is reached.
Testing - vSRL

- Images from imSitu dataset
- Activity classes from FrameNet
- Nouns from WordNet

The original dataset includes about 125,000 images with 75,702 for training, 25,200 for developing, and 25,200 for test. However, the dataset covers many non-human oriented activities (e.g., rearing, retrieving, and wagging), so we filter out these verbs, resulting in 212 verbs, leaving roughly 60,000 of the original 125,000 images in the dataset.

**Model:**

The model decomposes the probability of a realized situation, $y$, the combination of activity, $v$, and realized frame, a set of semantic (role,noun) pairs $(e, n_e)$, given an image $i$ as:

$$p(y|i; \theta) \propto \psi(v, i; \theta) \prod_{(e, n_e) \in R_f} \psi(v, e, n_e, i; \theta)$$

where each potential value in the CRF for subpart $x$, is computed using features $f_i$ from the VGG convolutional neural network (Simonyan and Zisserman, 2014) on an input image, as follows:

$$\psi(x, i; \theta) = e^{w^T f_i + b_x},$$

where $w$ and $b$ are the parameters of an affine transformation layer.
Testing - MCL

- Images from MS-COCO dataset

The dataset contains 80 object types but does not make gender distinctions between man and woman. We use the five associated image captions available for each image in this dataset to annotate the gender of people in the images. If any of the captions mention the word man or woman we mark it, removing any images that mention both genders. Finally, we filter any object category not strongly associated with humans by removing objects that do not occur with man or woman at least 100 times in the training set, leaving a total of 66 objects.

Model:

We decompose the joint probability of the output $y$, consisting of all object categories, $c$, and gender of the person, $g$, given an image $i$ as:

$$p(y|i; \theta) \propto \psi(g, i; \theta) \prod_{c \in y} \psi(g, c, i; \theta)$$

Where each potential value for $x$, is computed using features $f$, from a pertained ResNet-50 convolutional neural network, evaluated on the image,

$$\psi(x, i; \theta) = e^{w_x^T f_i + b_x}.$$

Trained a model using SGD with learning rate $10^{-5}$, momentum 0.9 and weight-decay $10^{-4}$, fine tuning the initial visual network, for 50 epochs.
Calibration

Calibration tries to enforce gender statistics derived from the training set of corpus applicable for each recognition problem.

**The inference problem for both models:**

\[
\arg \max_{y \in Y} f_\theta(y, i) = \log p(y| i; \theta).
\]

For all experiments, we try to match gender ratios on the test set within a margin of .05 of their value on the training set.

While we do adjust the output on the test set, we never use the ground truth on the test set and instead working from the assumption that it should be similarly distributed as the training set.

When running the debiasing algorithm, we set $\eta = 10^{-1}$ and optimize for 100 iterations.
Bias analysis

imSitu is gender biased:

Overall, the dataset is heavily biased toward male agents, with 64.6% of verbs favoring a male agent by an average bias of 0.707 (roughly 3:1 male). Nearly half of verbs are extremely biased in the male or female direction: 46.95% of verbs favor a gender with a bias of at least 0.7.

Training on imSitu amplifies bias:

The mean bias amplification in the development set is high, 0.050 on average, with 45.75% of verbs exhibiting amplification. Biased verbs tend to have stronger amplification: verbs with training bias over 0.7 in either the male or female direction have a mean amplification of 0.072.

MS-COCO is gender biased:

MS-COCO is even more heavily biased toward men than imSitu, with 86.6% of objects biased toward men, but with smaller average magnitude, 0.65. One third of the nouns are extremely biased toward males, 37.9% of nouns favor men with a bias of at least 0.7.

Training on MS-COCO amplifies bias:

The mean bias amplification across all objects is 0.036, with 65.67% of nouns exhibiting amplification. Larger training bias again tended to indicate higher bias amplification: biased objects with training bias over 0.7 had mean amplification of 0.081.
Calibration Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Viol.</th>
<th>Amp. bias</th>
<th>Perf. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>vSRL: Development Set</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRF</td>
<td>154</td>
<td>0.050</td>
<td>24.07</td>
</tr>
<tr>
<td>CRF + RBA</td>
<td>107</td>
<td>0.024</td>
<td>23.97</td>
</tr>
<tr>
<td>vSRL: Test Set</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRF</td>
<td>149</td>
<td>0.042</td>
<td>24.14</td>
</tr>
<tr>
<td>CRF + RBA</td>
<td>102</td>
<td>0.025</td>
<td>24.01</td>
</tr>
<tr>
<td>MLC: Development Set</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRF</td>
<td>40</td>
<td>0.032</td>
<td>45.27</td>
</tr>
<tr>
<td>CRF + RBA</td>
<td>24</td>
<td>0.022</td>
<td>45.19</td>
</tr>
<tr>
<td>MLC: Test Set</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRF</td>
<td>38</td>
<td>0.040</td>
<td>45.40</td>
</tr>
<tr>
<td>CRF + RBA</td>
<td>16</td>
<td>0.021</td>
<td>45.38</td>
</tr>
</tbody>
</table>

Table 2: Number of violated constraints, mean amplified bias, and test performance before and after calibration using RBA. The test performances of vSRL and MLC are measured by top-1 semantic role accuracy and top-1 mean average precision, respectively.

vSRL:

Development set:
The number of verbs whose bias exceed the original bias by over 5% decreases 30.5%.
Reduced bias amplification in vSRL by 52% on the development set.
Negligible decrease in performance.
Overall distance to the training set distribution after applying RBA decreased significantly, over 39%.

Test Set:
Decreased bias amplification by 40.5%.

MCL:

Development set:
Similarly to vSRL, we are able to reduce the number of objects whose bias exceeds the original training bias by 5%, by 40%.
Bias amplification was reduced by 31.3% on the development set.
Negligible decrease in performance.

Test Set:
Decreased bias amplification by 47.5%.
Figure 3: Results of reducing bias amplification using RBA on imSitu vSRL and MS-COCO MLC. Figures 3(a)-(d) show initial training set bias along the x-axis and development set bias along the y-axis. Dotted blue lines indicate the 0.05 margin used in RBA, with points violating the margin shown in red while points meeting the margin are shown in green. Across both settings adding RBA significantly reduces the number of violations, and reduces the bias amplification significantly. Figures 3(e)-(f) demonstrate bias amplification as a function of training bias, with and without RBA. Across all initial training biases, RBA is able to reduce the bias amplification.
Conclusion

- RBA does help to reduce bias

- General framework for visualizing and quantifying bias

- Can be extended to other social biases

- RBA can be applied to any structured predictor, it is unclear whether different predictors amplify bias more or less

- More extensive analysis could explore the interaction among predictor, bias measurement, and bias deamplification method

- Future work includes applying bias reducing methods in other structured domains, such as pronoun reference resolution (Mitkov, 2014)