MEN ALSO LIKE SHOPPING

REDUCING GENDER BIAS AMPLIFICATION USING CORPUS-LEVEL CONSTRAINTS

Jieyu Zhao\textsuperscript{1,3}, Tianlu Wang\textsuperscript{1}, \textbf{Mark Yatskar}\textsuperscript{2,4}, Vicente Ordonez\textsuperscript{1}, Kai-Wei Chang\textsuperscript{1,3}

\textsuperscript{1} University of Virginia \hspace{1cm} \textsuperscript{2} University of Washington \hspace{1cm} \textsuperscript{3} UCLA \hspace{1cm} \textsuperscript{4} Allen Institute for AI
Dataset Gender Bias

Male: 33%
Female: 66%

imsitu.org
Model Bias After Training

16\%  

84\%  

Male  

Female

imsitu.org
Why does this happen? Good for accuracy
Algorithmic Bias in Grounded Setting
Algorithmic Bias in Grounded Setting

World Dataset Model

woman cooking

cooking dusting faucet fork

World Dataset Model
Algorithmic Bias in Grounded Setting

- Woman cooking
- Man fixing faucet

World → Dataset → Model

{cooking dusting faucet fork}
Algorithmic Bias in Grounded Setting

Reduce amplification ~50%
Negligible loss in performance
Contributions

High dataset gender bias
38% (objects) 47% (events) exhibit strong bias

Models amplify existing gender bias
~70% objects and events have bias amplification

Reducing bias amplification
~50% reduction in amplification
Insignificant loss in performance

Data

Model

RBA
Outline

1. Background

2. Dataset Bias

3. Bias Amplification

4. Reducing Bias Amplification
imSitu Visual Semantic Role Labeling (vSRL) (events)

<table>
<thead>
<tr>
<th>ROLES</th>
<th>NOUNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGENT</td>
<td>woman</td>
</tr>
<tr>
<td>FOOD</td>
<td>vegetable</td>
</tr>
<tr>
<td>CONTAINER</td>
<td>pot</td>
</tr>
<tr>
<td>TOOL</td>
<td>spatula</td>
</tr>
</tbody>
</table>

FrameNet

WordNet

Internet

Yatskar et al. CVPR ’16, Yang et al. NAACL ’16, Gupta and Malik arXiv ’16
imSitu Visual Semantic Role Labeling (vSRL) (events)

<table>
<thead>
<tr>
<th>ROLES</th>
<th>NOUNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGENT</td>
<td>woman</td>
</tr>
<tr>
<td>FOOD</td>
<td>vegetable</td>
</tr>
<tr>
<td>CONTAINER</td>
<td>pot</td>
</tr>
<tr>
<td>TOOL</td>
<td>spatula</td>
</tr>
</tbody>
</table>
iMSitu Visual Semantic Role Labeling (vSRL) (events)

Convolutional Neural Network

Regression

Conditional Random Field

<table>
<thead>
<tr>
<th>ROLES</th>
<th>NOUNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGENT</td>
<td>woman</td>
</tr>
<tr>
<td>FOOD</td>
<td>vegetable</td>
</tr>
<tr>
<td>CONTAINER</td>
<td>pot</td>
</tr>
<tr>
<td>TOOL</td>
<td>spatula</td>
</tr>
</tbody>
</table>

Yatskar et al. CVPR ’16, Yang et al. NAACL ’16, Gupta and Malik arXiv ’16
**imSitu Visual Semantic Role Labeling (vSRL)**

(events)

- **ROLES**
  - **AGENT**: woman
  - **FOOD**: vegetable
  - **CONTAINER**: pot
  - **TOOL**: spatula

---

Yatskar et al. CVPR ’16, Yang et al. NAACL ’16, Gupta and Malik arXiv ’16
imSitu Visual Semantic Role Labeling (vSRL) (events)

Need to model correlation between variables
Model can use that machinery to amplify gender bias

<table>
<thead>
<tr>
<th>COOKING</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROLES</td>
</tr>
<tr>
<td>AGENT</td>
</tr>
<tr>
<td>FOOD</td>
</tr>
<tr>
<td>CONTAINER</td>
</tr>
<tr>
<td>TOOL</td>
</tr>
</tbody>
</table>

Convolutional Neural Network
Conditional Random Field
Regression

Yatskar et al. CVPR '16, Yang et al. NAACL '16, Gupta and Malik arXiv '16
a woman is smiling in a kitchen near a pizza on a stove

<table>
<thead>
<tr>
<th>WOMAN</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PIZZA</td>
<td>yes</td>
</tr>
<tr>
<td>ZEBRA</td>
<td>no</td>
</tr>
<tr>
<td>FRIDGE</td>
<td>yes</td>
</tr>
<tr>
<td>CAR</td>
<td>no</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>
COCO Multi-Label Classification (MLC) (objects)

| Woman | Pizza | Zebra | Fridge | Car | ...
|-------|-------|-------|--------|-----|------
|       | yes   | no    | yes    | no  |      
|       |       |       |        |     |      

Convolutional Neural Network

Regression

Conditional Random Field
Related Work

• Implicit Bias
  
  image search (Kay et al., 2015)
  search advertising (Sweeny, 2013)
  online news (Ross and Carter, 2011)
  credit score (Hardt et al., 2016)
  word vector (Bolukbasi et al., 2016)

• Classifier class imbalance
  
  Barocas and Selbst, 2014; Dwork et al., 2012;
  Feldman et al., 2015; Zliobaite, 2015
Outline

1. Background

2. Dataset Bias

3. Model Bias Amplification

4. Reducing Bias Amplification
Defining Dataset Bias (events)

Training Gender Ratio (verb)

Training Set

- cooking
- woman
- man

\[
\frac{\#(\text{cooking}, \text{man})}{\#(\text{cooking}, \text{man}) + \#(\text{cooking}, \text{woman})} = \frac{1}{3}
\]
Defining Dataset Bias (objects)

Training Gender Ratio (▲ noun)

Training Set

▲ snowboard

○ woman

● man

<table>
<thead>
<tr>
<th>MAN</th>
<th>snowboard</th>
<th>yes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>refrigerator</td>
<td>no</td>
</tr>
<tr>
<td></td>
<td>bowl</td>
<td>no</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>WOMAN</th>
<th>snowboard</th>
<th>yes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>refrigerator</td>
<td>no</td>
</tr>
<tr>
<td></td>
<td>bowl</td>
<td>no</td>
</tr>
</tbody>
</table>

\[
\frac{\#(▲ \text{snowboard}, \text{● man})}{\#(▲ \text{snowboard}, \text{● man}) + \#(▲ \text{snowboard}, \text{○ woman})} = \frac{2}{3}
\]
Gender Dataset Bias

- **imSitu Verb**
- **COCO Noun**

Graph showing gender bias in dataset with X-axis representing unbiased gender ratio and Y-axis representing % of items. Two lines indicate bias towards male (orange) and female (red).
Gender Dataset Bias

- ImSitu Verb
- COCO Noun

% of items

Gender Ratio

Female bias

Male bias

Unbiased

Gender Ratio

Gender Ratio

- Shopping
- Braiding
- Washing
- Cooking
- Lecturing
- Coaching
- Repairing
Gender Dataset Bias

- **imSitu Verb**
  - 64.6% bias
  - 46.9% strong bias (>2:1)

- **COCO Noun**
  - 86.6% bias
  - 37.9% strong bias (>2:1)

% of items

Unbiased Gender Ratio

Female bias

Male bias
Outline

1. Background

2. Dataset Bias

3. Bias Amplification

4. Reducing Bias Amplification

ImSitu vSRL (events)

COCO MLC (objects)
Defining Bias Amplification (events)
Predicted Gender Ratio (verb)

Development Set

What does the model predict on unseen data?
Defining Bias Amplification (events)

Predicted Gender Ratio (verb)

Development Set

- cooking
- woman
- man

\[
\frac{\#(\text{cooking}, \text{man})}{\#(\text{cooking}, \text{man}) + \#(\text{cooking}, \text{woman})} = \frac{1}{6}
\]
Model Bias Amplification

- imSitu Verb
- COCO Noun

Predicted Gender Ratio

Imbalanced Gender Ratio

Unbiased Gender Ratio

Amplification Zone

Amplification

Unmatched gender ratio

Female bias

Male bias
Model Bias Amplification

- **imSitu Verb**
- **COCO Noun**

- Predicted Gender Ratio
  - 0.00
  - 0.25
  - 0.50
  - 0.75
  - 1.00

- Gender Ratio
  - Unbiased
  - Male bias
  - Female bias

- Matched gender ratio

- Amplification Zone
- Amplification Zone

- Washing
- Cooking
- Autographing
- Assembling

- Female bias
- Male bias

- Unbiased Gender Ratio
  - 0.00
  - 0.25
  - 0.50
  - 0.75
  - 1.00

- Matched gender ratio

- Model Bias Amplification

- Graph with data points representing imSitu verbs and COCO nouns.
Model Bias Amplification

- **imSitu Verb**: 69% bias↑ .05 |bias↑| > 2:1 initial bias : .07 |bias↑|
- **COCO Noun**: 73% bias↑ .04 |bias↑| > 2:1 initial bias : .08 |bias↑|

---

**Predicted Gender Ratio**

**Unbiased Gender Ratio**

- Female bias
- Male bias

**Amplification Zone**

- Yellow
- Blue line: Matched gender ratio

---

35
Summary

Can we remove gender bias amplification and still maintain performance?
Can we remove gender bias amplification and still maintain performance?

**Performance Goal:** as good as the original

**Fairness Goal:** not more biased than the data it was trained on
Outline

1. Background

2. Dataset Bias

3. Bias Amplification

4. Reducing Bias Amplification
Reducing Bias Amplification (RBA)

- Corpus level constraints on model output (ILP)
  - Doesn’t require model retraining
- Reuse model inference through Lagrangian relaxation
  - Can be applied to any structured model

- Doesn’t require model retraining
Reducing Bias Amplification (RBA)

Integer Linear Program

\[
\sum_i \max_{y_i} s(y_i, \text{image}) \quad \rightarrow \quad \text{base model}
\]

\[
\text{CRF Inference}
\]
Reducing Bias Amplification (RBA)

Integer Linear Program

\[
\sum_i \max y_i \cdot s(y_i, \text{image})
\]

\forall \text{ points } \left| \frac{\text{Training Ratio} - \text{Predicted Ratio}}{f(y_1 \ldots y_n)} \right| \leq \text{margin}

![Graph showing the relationship between predicted gender ratio and actual gender ratio, with points indicating whether they are within or violating the margin.](image)
Reducing Bias Amplification (RBA)

Integer Linear Program

\[ \sum_{i} \max_{y_i} \ s(y_i, \text{image}) \]

\[ \forall \text{ points} \quad \left| \frac{\text{Training Ratio} - \text{Predicted Ratio}}{f(y_1 \ldots y_n)} \right| \leq \text{margin} \]
Reducing Bias Amplification (RBA)

Integer Linear Program

$$\sum \max_{\gamma_i} s(\gamma_i, \text{image})$$

$$\forall \text{ points} \quad | \text{Training Ratio} - \text{Predicted Ratio} | \leq \text{margin}$$

\[ f(\gamma_1 \ldots \gamma_n) \]

Lagrangian Relaxation

\[ \text{inference} \quad \text{constraints} \]

Sontag et al., 2011; Rush and Collins, 2012; Chang and Collins, 2011; Peng et al., 2015, Chang et al., 2013; Dalvi, 2015
Lagrangian Relaxation

\[ \sum \max_i s(y_i, \text{image}) \]

\[ \left| \frac{\text{Training Ratio} - \text{Predicted Ratio}}{} \right| \leq \text{margin} \]

\(| (1/2) |
Lagrangian Relaxation

\[ \sum_{i} \max_{y_i} s(y_i, \text{image}) \]

\[ \left| \text{Training Ratio - Predicted Ratio} \right| \leq \text{margin} \quad (1/2) \]

- Lagrange Multiplier (\( \lambda \)) Per Constraint

- Inference
- Update \( \lambda \)
- Update potentials

Sontag et al., 2011; Rush and Collins, 2012; Chang and Collins, 2011; Peng et al., 2015, Chang et al., 2013; Dalvi, 2015
Lagrangian Relaxation

\[
\sum \max_i s(y_i, \text{image}) \\
\left| \frac{\text{Training Ratio} - \text{Predicted Ratio}}{2} \right| \leq \text{margin}
\]

- Lagrange Multiplier ($\lambda$) Per Constraint

- Inference
- Update $\lambda$
- Update potentials

Sontag et al., 2011; Rush and Collins, 2012; Chang and Collins, 2011; Peng et al., 2015; Chang et al., 2013; Dalvi, 2015
Lagrangian Relaxation

\[ \sum \max_{i} s(y_i, \text{image}) \]

\[ \left| \frac{\text{Training Ratio} - \text{Predicted Ratio}}{\text{margin}} \right| \leq \frac{1}{2} \]

- Lagrange Multiplier (\(\lambda\)) Per Constraint

\[ \text{inference} \]

\[ \text{update } \lambda \]

\[ \text{update potentials} \]

Sontag et al., 2011; Rush and Collins, 2012; Chang and Collins, 2011; Peng et al., 2015, Chang et al., 2013; Dalvi, 2015
Lagrangian Relaxation

\[ \sum \max_i s(y_i, \text{image}) \]

\[ \gamma_i \]

Training Ratio - Predicted Ratio \[ \leq \text{margin} \]

(1/2)

- Lagrange Multiplier (\( \lambda \)) Per Constraint

- **inference**

- **update \( \lambda \)**

- **update potentials**

Sontag et al., 2011; Rush and Collins, 2012; Chang and Collins, 2011; Peng et al., 2015; Chang et al., 2013; Dalvi, 2015
Lagrangian Relaxation

\[
\sum_i \text{max } s(y_i, \text{image})
\]

Training Ratio - Predicted Ratio \(\leq \text{margin (1/2)}\)

- Lagrange Multiplier (\(\lambda\)) Per Constraint

inference

update \(\lambda\)

update potentials

Sontag et al., 2011; Rush and Collins, 2012; Chang and Collins, 2011; Peng et al., 2015, Chang et al., 2013; Dalvi, 2015
Lagrangian Relaxation

\[
\sum_i \max \ s(y_i, \text{image}) \\
\left| \frac{\text{Training Ratio}}{\text{Predicted Ratio}} \right| \leq \text{margin} (1/2)
\]

- Lagrange Multiplier ($\lambda$) Per Constraint

- Inference
- Update $\lambda$
- Update potentials
Gender Bias De-amplification in imSitu

imSitu Verb  Violation:  72.6%  .050 |bias↑|  24.07 acc.
Gender Bias De-amplification in imSitu

<table>
<thead>
<tr>
<th>imSitu Verb</th>
<th>Violation</th>
<th></th>
<th>bias↑</th>
<th>acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/ RBA</td>
<td>50.5%</td>
<td>.024</td>
<td>bias↑</td>
<td>23.97 acc.</td>
</tr>
<tr>
<td></td>
<td>72.6%</td>
<td>.050</td>
<td>bias↑</td>
<td>24.07 acc.</td>
</tr>
</tbody>
</table>

Predicted Gender Ratio

- Violating margin
- Within margin
- Margin
- Matched gender ratio

Female bias

Male bias

Unbiased Gender Ratio

52
Gender Bias De-amplification in COCO

COCO Noun  Violation: 60.6%  \( .032 \text{ |bias|} \)  45.27 mAP

Predicted Gender Ratio

Unbiased Gender Ratio

- Violating margin
- Within margin
- Margin
- Matched gender ratio

Female bias  Male bias
Gender Bias De-amplification in COCO

<table>
<thead>
<tr>
<th>COCO Noun</th>
<th>Violation: 60.6%</th>
<th>Bias↑</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/ RBA</td>
<td>36.4%</td>
<td>.022</td>
<td>45.19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predicted Gender Ratio</th>
<th>0</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female bias</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>0.75</td>
<td>1</td>
</tr>
<tr>
<td>Male bias</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>0.75</td>
<td>1</td>
</tr>
</tbody>
</table>

- Violating margin
- Within margin
- Margin
- Matched gender ratio

Predicted Gender Ratio vs. Unbiased Gender Ratio

- Red dots: Violating margin
- Green dots: Within margin
Gender Bias De-amplification in COCO

COCO Noun  
Violation: 60.6%  0.032 \text{bias}  45.27 \text{mAP}

w/ RBA  
Violation: 36.4%  0.022 \text{bias}  45.19 \text{mAP}

Performance Goal: as good as the original  
Fairness Goal: not more biased than the data it was trained on
Contributions

- **High dataset gender bias**: 38% (objects) and 47% (events) exhibit strong bias.
- **Models amplify existing gender bias**: ~70% objects and events have bias amplification.
Future Work

Can existing data be made more balanced?

Do all models amplify equally?
  i.e. different objectives

Other direct applications?
  i.e. co-ref, racial bias
Questions?

https://github.com/uclanlp/reducingbias

imSitu vSRL (events)

COCO MLC (objects)