Mitigating Gender Bias Amplification in Distribution by Posterior Regularization

Shengyu Jia[♦*], **Tao Meng**[♣*], Jieyu Zhao[♣], Kai-Wei Chang[♣]

[♦]Tsinghua University

[♣]University of California, Los Angeles

Dataset Gender Bias

33%

Male

66%

Female

imsitu.org

Credit to Mark Yatskar

Prediction Gender Bias

16%

Male

84%

Female

imsitu.org

Credit to Mark Yatskar

Top Prediction vs. Distribution Prediction

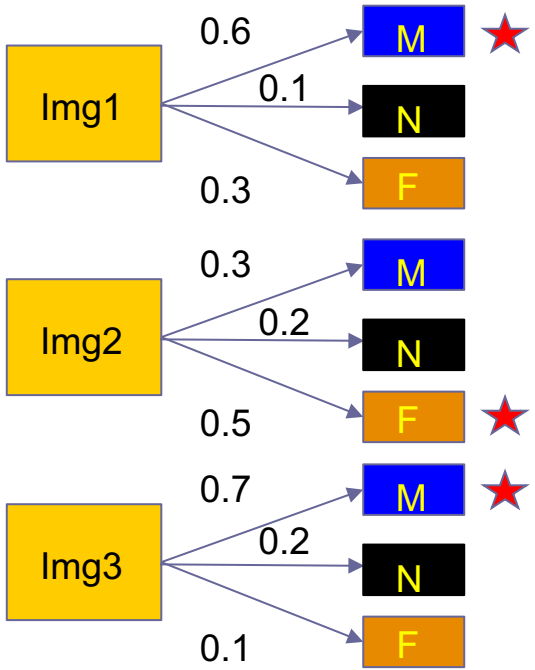
- Visual Semantic Role Labelling (vSRL)
  - CNN: Feature extraction
  - CRF: Assign every instance a probability

- Top prediction (Zhao et. al. 17):
  - Model is forced to make one decision
    - Even similar probabilities for “female” and “male” predictions
  - Potentially amplify the bias

- ※Distribution of predictions (this work):
  - A better view of understanding bias amplification
  - Model is trained using regularized maximum likelihood objective

Bias Amplification in Distribution

Bias in top predictions (Zhao et. al. 17):

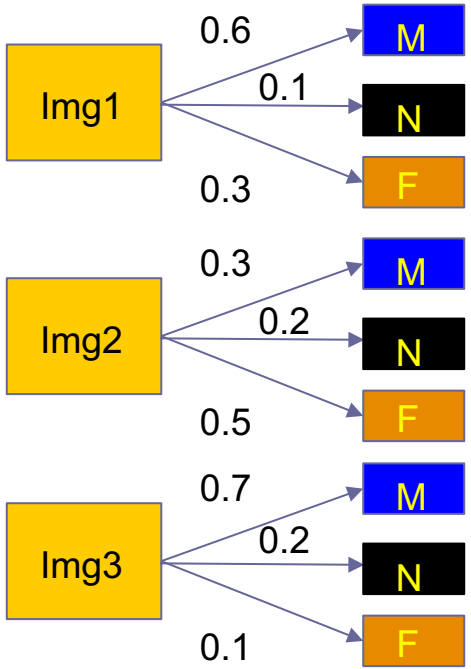


Towards Male:

bias_pred = $\frac{M \quad M}{M \quad F \quad M}$ = 0.67

Bias Amplification in Distribution

Bias in posterior distribution:

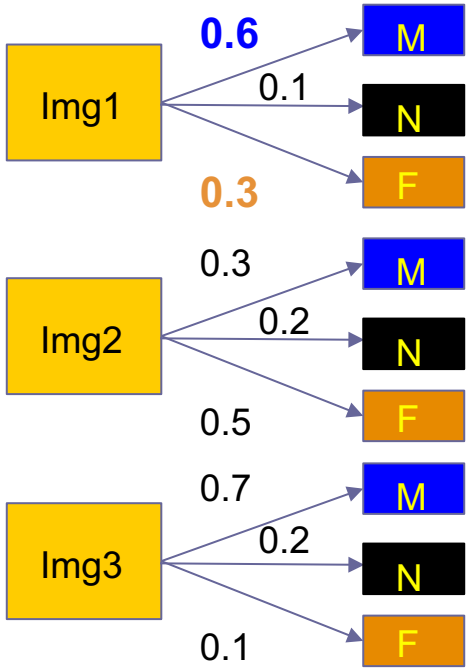


Towards Male:

bias_dist =

Bias Amplification in Distribution

Bias in posterior distribution:

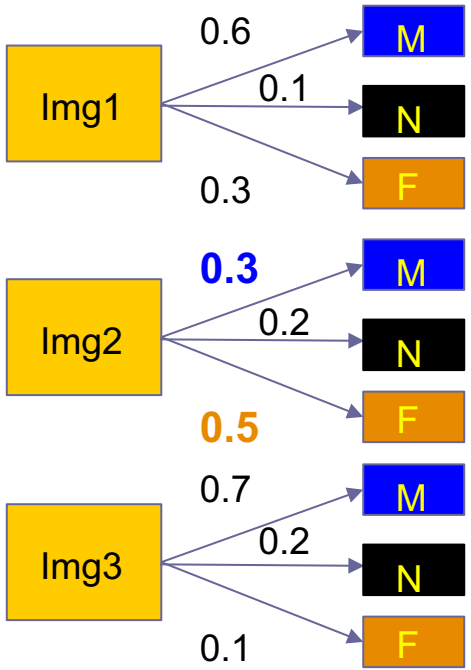


Towards Male:

$$\text{bias\_dist} = \frac{0.6}{(0.6 + 0.3)}$$

Bias Amplification in Distribution

Bias in posterior distribution:

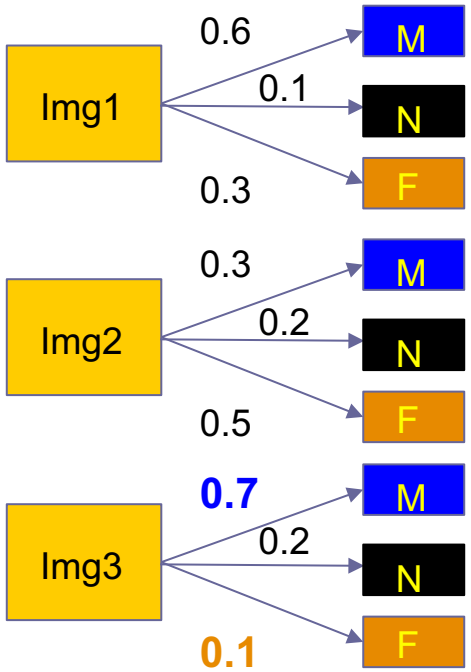


Towards Male:

$$\text{bias\_dist} = \frac{0.6 + 0.3}{(0.6 + 0.3) + (0.3 + 0.5)}$$

Bias Amplification in Distribution

Bias in posterior distribution:

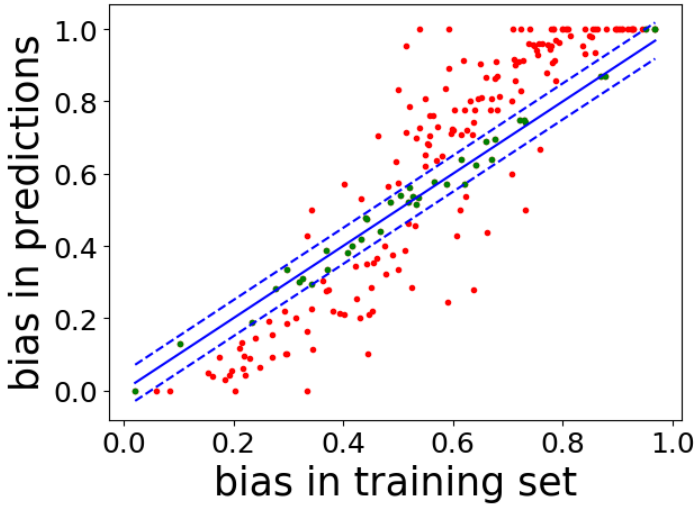


Towards Male:

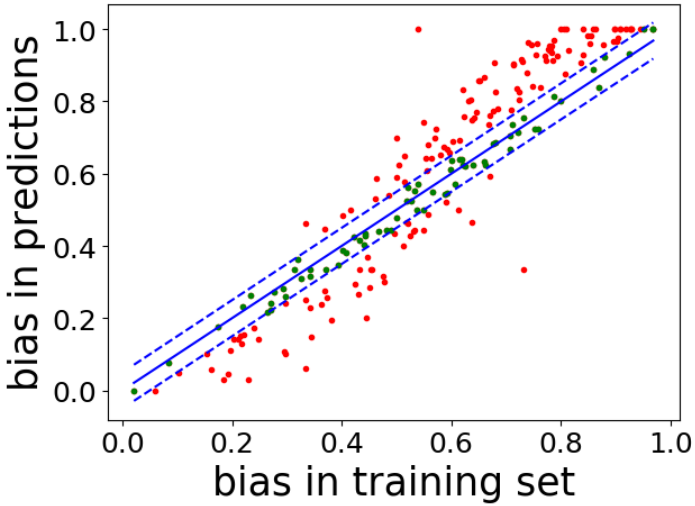
$$\text{bias\_dist} = \frac{0.6 + 0.3 + 0.7}{(0.6 + 0.3) + (0.3 + 0.5) + (0.7 + 0.1)} = 0.59$$

Bias Amplification in Distribution

- In top predictions the bias is amplified (left, 81.6% violations).
- Similar to top predictions, the posterior distribution perspective also indicates bias amplification. (right, 51.4% violations)



Top prediction (Zhao et. al. 17)



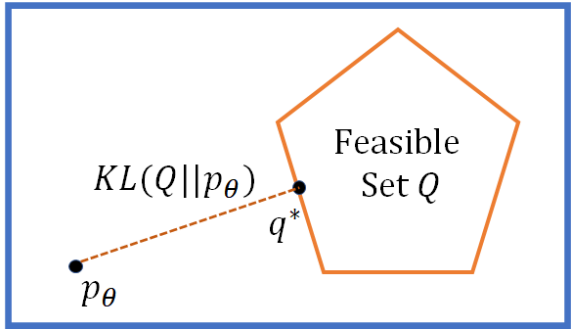
Posterior Distribution

Posterior Regularization (PR) for Mitigation

1. Define the constraints and the feasible set Q:
   the posterior bias should be similar to the bias in the training set.

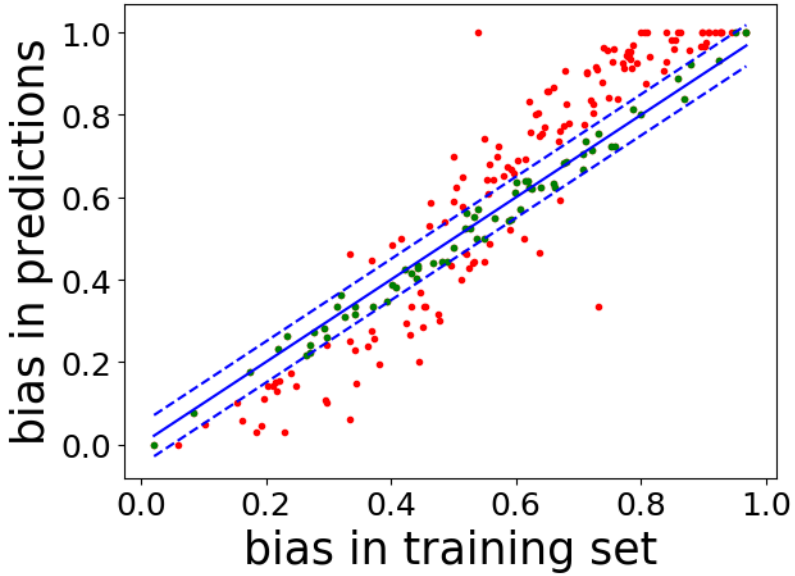
1. Minimize the KL-divergence

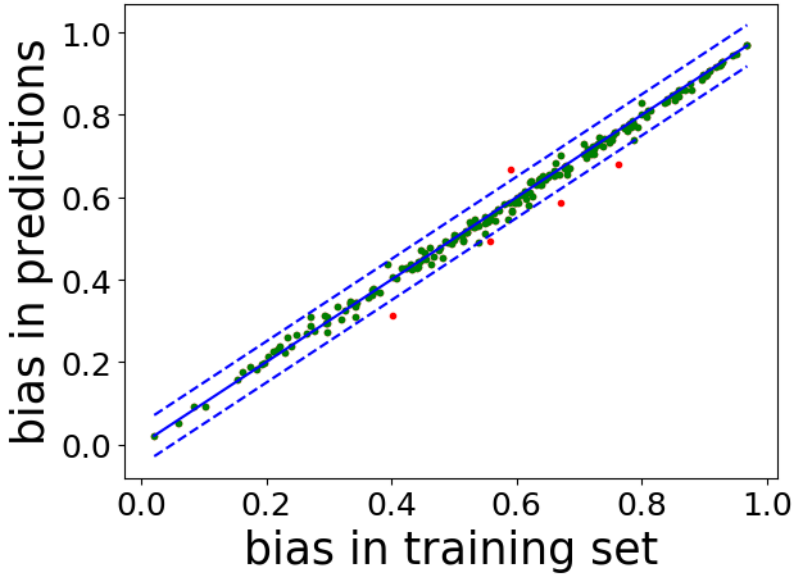
1. Do MAP inference based on the regularized distribution



Amplification Mitigation Using PR

vSRL	Violation: 51.4%	Amplification: 0.032	Accuracy: 23.2%
w/ PR	Violation: 2%	Amplification: -0.005	Accuracy: 23.1%





Conclusion

1. Analyze bias amplification from distribution perspectiv。 e

2. Remove almost all the bias amplification using PR。

3. Open question: why the bias in posterior distribution is amplified.

https://github.com/uclanlp/reducingbias