SONY

ICML
International Conference On Machine Learning

Men Also Do Laundry: Multi-Attribute Bias Amplification

Dora Zhao[1], Jerone T. A. Andrews[2], and Alice Xiang[1]

[1]Sony AI, New York, [2]Sony AI, Tokyo

Sony AI

Our contributions

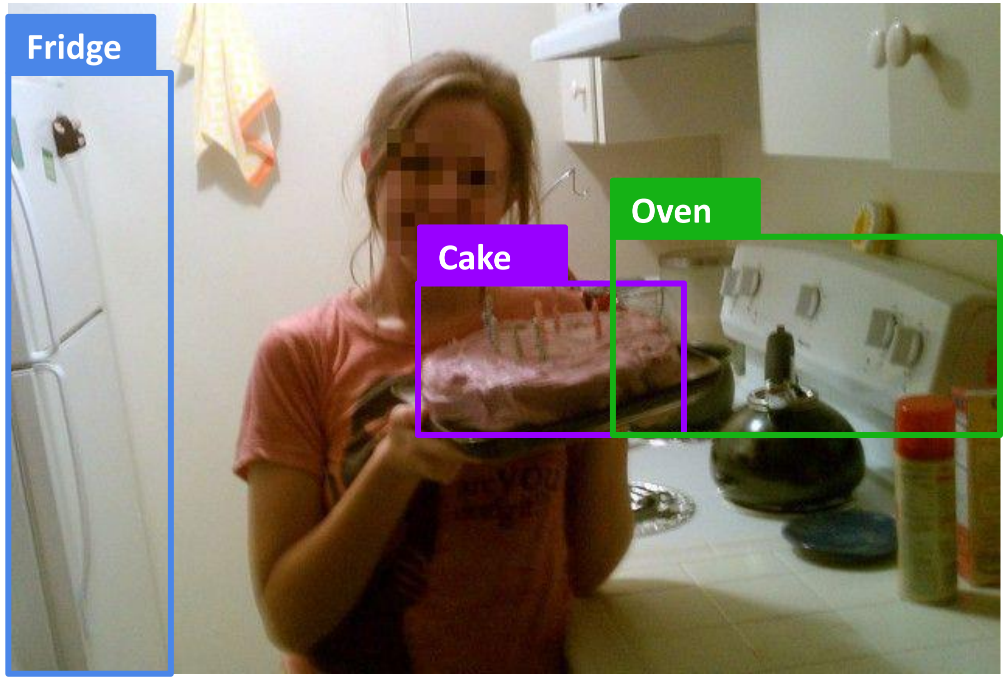
Contributions.

1. Present two new metrics that measure bias amplification with respect to multiple attributes
2. Empirically evaluate the presence of multi-attribute bias amplification on three datasets: COCO [1], imSitu [2], and CelebA [3]
3. Present a novel evaluation of bias amplification on a non-binary group (`hair color` in CelebA)
4. Benchmark existing bias mitigation techniques [4, 5, 6, 7] using single and multi-attribute bias amplification metrics

[1] Lin et al. "Microsoft COCO: Common Objects in Context." ECCV 2014.
[2] Yatskar et al. "Situation Recognition: Visual Semantic Role Labeling for Image Understanding." CVPR 2016.
[3] Liu et al. "Deep Learning Face Attributes in the Wild." ICCV 2015.
[4] Zhao et al. "Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints." EMNLP 2017.
[5] Wang et al. "Balanced Datasets are Not Enough: Estimating and Mitigating Gender Bias in Deep image Representations." ICCV 2019.
[6] Wang et al. "Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation." CVPR 2020.
[7] Agarwal et al. "Does Data Repair Lead to Fair Models? Curating Contextually Fair Data to Reduce Model Bias." WACV 2022.

Existing methods measure single-attribute bias amplification

- **Bias amplification** occurs when a model compounds the inherent biases of its training set at test time [1]
- There are two main approaches to measuring bias amplification in computer vision:
    1. Leakage-based metrics [2, 3]
    2. Co-occurrence-based metrics [1, 4]
- However, most of these approaches measure bias amplification wrt single annotated attributes (e.g., `fridge`). However, most images in computer vision datasets contain **multiple attribute annotations** (e.g., {`fridge, oven, cake`})



[1] Zhao et al. "Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints." EMNLP 2017.
[2] Wang et al. "Balanced Datasets are Not Enough: Estimating and Mitigating Gender Bias in Deep image Representations." ICCV 2019.
[3] Hirota et al. "Quantifying Societal Bias Amplification in Image Captioning." CVPR 2022.
[4] Wang and Russakovsky. "Directional Bias Amplification." ICML 2021.

Multi-attribute bias amplification

We propose **two** co-occurrence-based metric that takes into account multiple attributes:

1) **Undirected Bias Amplification**

$$\text{Multi}_{\text{MALS}} = X, \text{Var}(\Delta_{mg}) \quad \text{where}$$

Set of all attribute combinations

$$X = \frac{1}{|\mathcal{M}|} \sum_{g \in \mathcal{G}} \sum_{m \in \mathcal{M}} |\Delta_{gm}|$$

$$\Delta_{gm} = 1\left[\text{bias}_{\text{train}}(m, g) > |\mathcal{G}|^{-1}\right] \cdot \left(\text{bias}_{\text{test}}(m, g) - \text{bias}_{\text{train}}(m, g)\right)$$

Set of group membership labels

Co-occurrence ratio between attribute combination and group

2) **Directed Bias Amplification**

$$\text{Multi}_{\rightarrow} = X, \text{Var}(\Delta_{mg}) \quad \text{where}$$

$$X = \frac{1}{|\mathcal{G}||\mathcal{M}|} \sum_{g \in \mathcal{G}} \sum_{m \in \mathcal{M}} y_{gm} |\Delta_{gm}| + (1 - y_{gm}) |-\Delta_{gm}| \,,$$

$$y_{gm} = 1[P_{\text{train}}(g = 1, m = 1) > P_{\text{train}}(g = 1) P_{\text{train}}(m = 1)]$$

$$\Delta_{gm} = \begin{cases} P_{\text{test}}(\hat{m} = 1 | g = 1) - P_{\text{train}}(m = 1 | g = 1) & \text{if measuring } G \rightarrow M \\ P_{\text{test}}(\hat{g} = 1 | m = 1) - P_{\text{train}}(g = 1 | m = 1) & \text{if measuring } M \rightarrow G \end{cases}$$

Advantages

1. Our metric accounts for co-occurrences with multiple attributes

2. Negative and positive values do not cancel each other out

3. Our metric is more interpretable

Evaluating bias amplification on existing computer vision datasets

We benchmark our metric and existing single-attribute bias amplification metrics [1, 2] using three datasets:

Dataset	Group	Attribute
COCO [3]	Perceived gender expression `{female, male}`	52 objects
imSitu [4]	Perceived gender expression `{female, male}`	Action, location
CelebA [5]	Hair color `{blonde hair, black hair, brown hair}`	23 physical attributes

[1] Zhao et al. “Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints.” EMNLP 2017.
[2] Wang and Russakovsky. “Directional Bias Amplification.” ICML 2021.
[3] Lin et al. “Microsoft COCO: Common Objects in Context.” ECCV 2014.
[4] Yatskar et al. “Situation Recognition: Visual Semantic Role Labeling for Image Understanding.” CVPR 2016.
[5] Liu et al. “Deep Learning Face Attributes in the Wild.” *ICCV 2015.*

Bias amplification from multiple attributes is greater than from single attributes

To analyze the effect of considering multiple attributes, we perform evaluation on datasets that are balanced w.r.t. single attributes.

(a) COCO	$\|m_i\| \geq 2$	$\|m_i\| \geq 1$
$\text{Multi}_{\text{MALS}}$	22.3 ± 0.7, $[4.6 \pm 0.1]$	21.9 ± 0.2, $[4.5 \pm 0.1]$
$\text{Multi}_{M \to G}$	22.7 ± 0.3, $[12.9 \pm 0.2]$	22.2 ± 0.3, $[13.0 \pm 0.0]$
$\text{Multi}_{G \to M}$	0.3 ± 0.0, $[0.0 \pm 0.0]$	0.3 ± 0.0, $[0.0 \pm 0.0]$
(b) imSitu	$\|m_i\| \geq 2$	$\|m_i\| \geq 1$
$\text{Multi}_{\text{MALS}}$	18.0 ± 0.3, $[3.0 \pm 0.1]$	9.4 ± 0.2, $[1.6 \pm 0.1]$
$\text{Multi}_{M \to G}$	14.5 ± 0.2, $[4.1 \pm 0.2]$	13.0 ± 0.1, $[3.2 \pm 0.1]$
$\text{Multi}_{G \to M}$	0.1 ± 0.0, $[0.0 \pm 0.0]$	0.1 ± 0.0, $[0.0 \pm 0.0]$
(c) CelebA	$\|m_i\| \geq 2$	$\|m_i\| \geq 1$
$\text{Multi}_{\text{MALS}}$	23.2 ± 0.4, $[2.3 \pm 0.1]$	23.1 ± 0.4, $[2.3 \pm 0.1]$
$\text{Multi}_{M \to G}$	5.5 ± 0.0, $[0.0 \pm 0.0]$	5.5 ± 0.0, $[0.0 \pm .0.0]$
$\text{Multi}_{G \to M}$	0.6 ± 0.0, $[0.1 \pm 0.0]$	0.6 ± 0.0, $[0.1 \pm 0.0]$

- We show multi-attribute bias amplification (mean and variance) when varying $|m_i|$, the minimum number of attributes in a combination.
- $\text{Multi}_{\text{MALS}}$ increases for $|m_i| \geq 2$ compared to $|m_i| \geq 1$

Single-attribute bias amplification methods can increase multi-attribute amplification

We benchmark five bias mitigation methods [1, 2, 3, 4] trained and evaluated using the unbalanced dataset.

(a) COCO	mAP	BiasAmp$_{\text{MALS}}$	Multi$_{\text{MALS}}$	BiasAmp$_{M\to G}$	Multi$_{M\to G}$	BiasAmp$_{G\to M}$	Multi$_{G\to M}$
Original	53.4 ± 0.2	-0.6 ± 0.3	14.5 ± 0.6	2.2 ± 0.4	12.5 ± 0.2	$\mathbf{-0.0 \pm 0.0}$	0.4 ± 0.0
Oversampling	51.5 ± 0.1	1.1 ± 0.1	14.0 ± 0.4	-3.4 ± 0.2	12.5 ± 0.3	-0.2 ± 0.0	$\mathbf{0.3 \pm 0.0}$
RBA	50.7 ± 1.1	3.8 ± 1.7	14.9 ± 1.1	-6.3 ± 3.5	17.3 ± 2.2	0.1 ± 01	0.4 ± 0.0
Adv	$\mathbf{59.0 \pm 0.1}$	-0.7 ± 0.9	17.1 ± 0.4	7.0 ± 0.6	14.7 ± 0.6	0.1 ± 0.0	0.3 ± 0.0
DomInd	56.1 ± 0.3	0.4 ± 0.6	$\mathbf{12.6 \pm 0.8}$	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$	0.3 ± 0.0	0.3 ± 0.0
Data Repair	48.5 ± 0.1	$\mathbf{0.3 \pm 0.1}$	17.2 ± 0.3	1.9 ± 0.3	11.7 ± 0.2	-0.0 ± 0.0	0.4 ± 0.0
(b) imSitu	**mAP**	**BiasAmp$_{\text{MALS}}$**	**Multi$_{\text{MALS}}$**	**BiasAmp$_{M\to G}$**	**Multi$_{M\to G}$**	**BiasAmp$_{G\to M}$**	**Multi$_{G\to M}$**
Original	67.1 ± 0.1	2.5 ± 0.1	37.5 ± 0.1	-0.3 ± 0.1	20.6 ± 0.1	0.0 ± 0.0	0.2 ± 0.0
Oversampling	66.3 ± 0.1	-4.5 ± 0.2	35.8 ± 0.1	-2.4 ± 0.1	20.1 ± 0.1	-0.0 ± 0.0	0.2 ± 0.0
RBA	54.7 ± 0.5	$\mathbf{-1.4 \pm 0.3}$	35.4 ± 0.3	-6.2 ± 0.3	40.7 ± 0.5	-0.1 ± 0.0	0.3 ± 0.0
Adv	58.1 ± 0.1	4.1 ± 0.3	38.7 ± 0.3	0.6 ± 0.4	28.1 ± 0.3	-0.0 ± 0.0	0.2 ± 0.0
DomInd	$\mathbf{69.6 \pm 0.1}$	10.2 ± 0.9	37.5 ± 0.4	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$	0.1 ± 0.0	0.2 ± 0.0
Data Repair	62.3 ± 0.1	-1.8 ± 0.1	$\mathbf{16.2 \pm 0.1}$	-0.1 ± 0.1	24.2 ± 0.1	$\mathbf{-0.0 \pm 0.0}$	$\mathbf{0.1 \pm 0.0}$

[1] Zhao et al. "Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints." EMNLP 2017.
[2] Wang et al. "Balanced Datasets are Not Enough: Estimating and Mitigating Gender Bias in Deep image Representations." ICCV 2019.
[3] Wang et al. "Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation." CVPR 2020.
[4] Agarwal et al. "Does Data Repair Lead to Fair Models? Curating Contextually Fair Data to Reduce Model Bias." WACV 2022.

Key takeaways

1. Models can leverage correlations between groups and multiple attributes simultaneously
2. On average, bias amplification from multiple attributes is greater than that from single attributes
3. Single attribute bias mitigation methods can inadvertently increase multi-bias amplification

Paper

ICML Page

SONY