Bias in Language

Eve Fleisig and Liwei Song
April 9, 2020

Real-World Problems

Google accused of racism after black names are 25% more likely to bring up adverts for criminal records checks

- Professor finds 'significant discrimination' in ad results, with black names 25 per cent more likely to be linked to arrest record check services

Amazon apologises for 'ham-fisted' error that made gay books 'disappear'

Firm apologises for sales ranking system mistake that hit books dealing with gay themes

Amazon scraps secret AI recruiting tool that showed bias against women

Motivation

- Biases in NLP models perpetuate stereotypes
- Stereotype-based biases worsen model performance

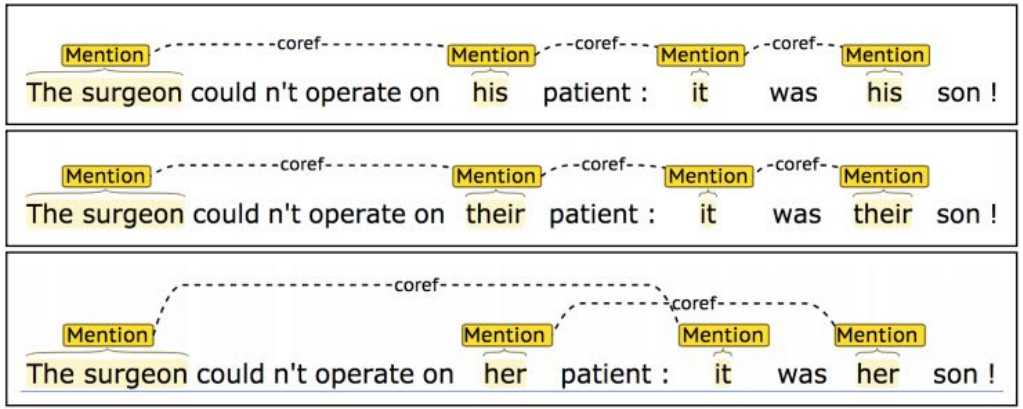
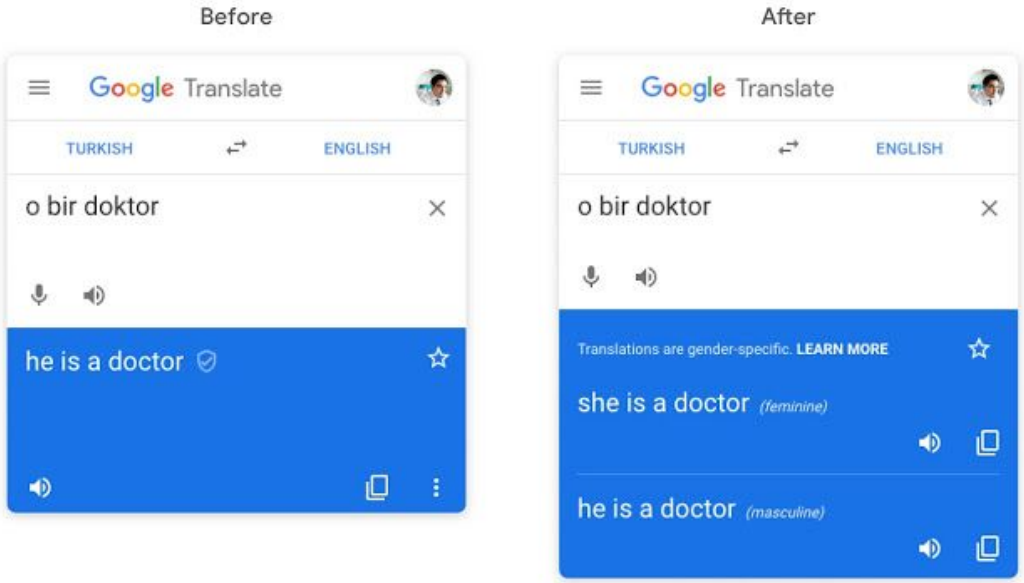


Figure 1: Stanford CoreNLP rule-based coreference system resolves a male and neutral pronoun as coreferent with "The surgeon," but does not for the corresponding female pronoun.

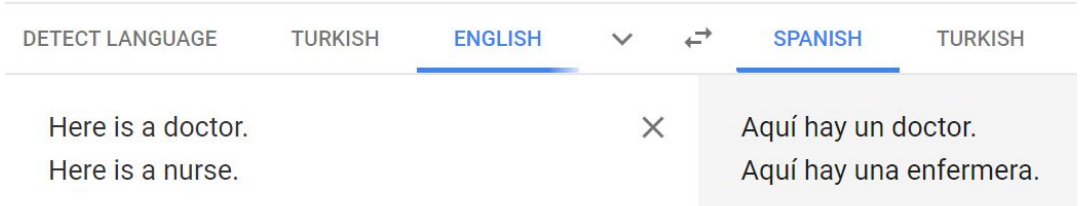
Rudinger et al. (2018)

Motivation

- Despite some efforts, biases persist today



https://ai.googleblog.com/2018/12/providing-gender-specific-translations.html



Overview

Measuring Bias:

Semantics derived automatically from language corpora contain human-like biases (Caliskan et al., *Science 2017*)

On Measuring Social Biases in Sentence Encoders (May et al., *NAACL 2019*)

Reducing Bias:

Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints (Zhao et al., *EMNLP 2017*; best long paper award)

The Bigger Picture

Overview

Measuring Bias:

Semantics derived automatically from language corpora contain human-like biases (Caliskan et al., *Science 2017*)

On Measuring Social Biases in Sentence Encoders (May et al., *NAACL 2019*)

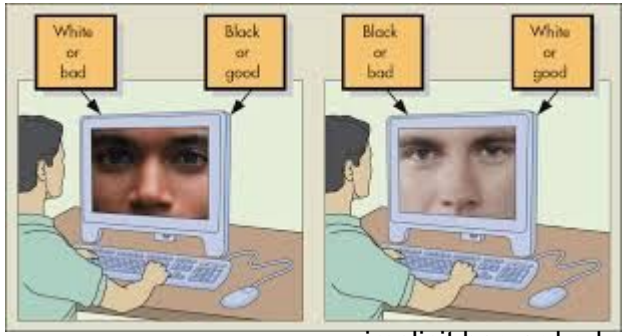
Reducing Bias:

Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints (Zhao et al., EMNLP 2017; best long paper award)

The Bigger Picture

Word Embedding Association Test (Caliskan et al., 2017)

- Measure bias in word embeddings (GloVe and word2vec)
- Based on Implicit Association Test
- Measure association between **target words** and **attribute words**



implicit.harvard.edu

Target Words	
X ("European American Names")	Y ("African American Names")
Adam, Harry, Nancy...	Jamel, Lavar, Latisha...

Attribute Words	
A ("Pleasant Attributes")	B ("Unpleasant Attributes")
love, cheer, friend...	ugly, evil, abuse...

Word Embedding Association Test

- Difference between sums of s(w, A, B), where s(w, A, B) is the difference in mean **cosine similarities**

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

$$s(w, A, B) = \text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b)$$

Word Embedding Association Test (WEAT)

- Effect size: Measures strength of association

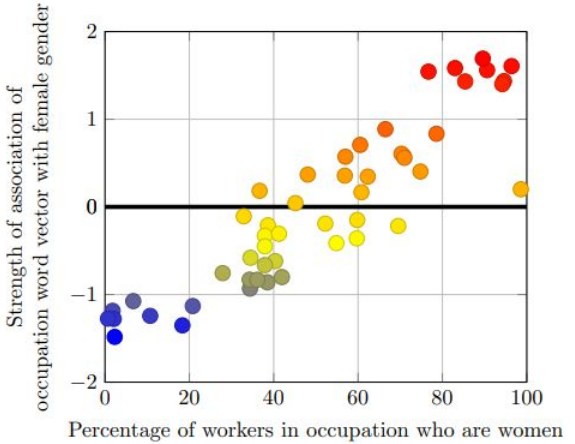
$$d = \frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{stddev}_{w \in X \cup Y} s(w, A, B)}$$

- Permutation test to compute significance:

$$p = \Pr[s(X_i, Y_i, A, B) > s(X, Y, A, B)]$$

Tests and Results

- Significant biases for several tests
- Word embeddings for professions correlate strongly with % of women in that profession
- Stereotypes reflected in statistical machine translation



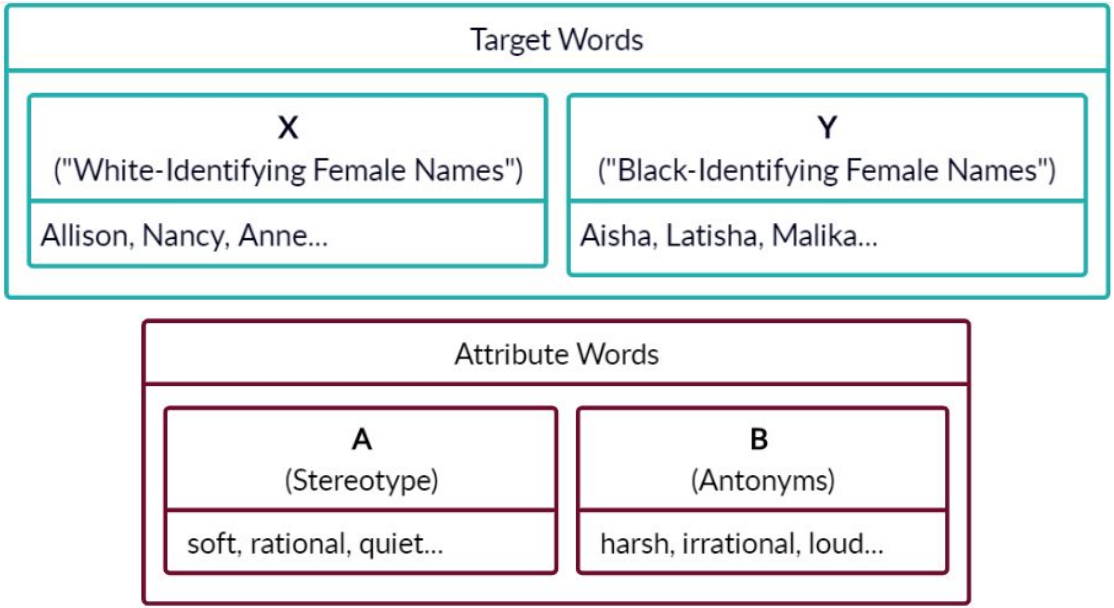
		Human IAT		WEAT	
Target Words	Attribute Words	d (effect size)	p-value	d (effect size)	p-value
Flowers/insects	Pleasant/unpleasant	1.35	10^{-8}	1.54	10^{-7}
European-/African-American names	Pleasant/unpleasant	N/A	N/A	1.24*	10^{-3}
Male/female names	Career/family	0.72	10^{-2}	1.89	10^{-4}
Math/arts	Male/female terms	0.82	$<10^{-2}$	0.97	.027
Science/arts	Male/female terms	1.47	10^{-24}	1.24	10^{-2}

Sentence Encoder Association Test (May et al., 2019)

- Extend WEAT to sentence encoders: same formulae
- Slot each word into a semantically bleached sentence template
  - Minimize effect of context
  - Focus on what attributes the encoder associates with a target word
- Apply WEAT to the vector embeddings of the target and attribute sentences

Biases Tested: Angry Black Woman Stereotype

- Black women portrayed as loud, imposing
- Target words: white- and black-identifying female names
- Attribute words: Words associated with stereotype vs. their antonyms



Biases Tested: Double Bind

- Women who clearly succeed are perceived as less likable
- Women who ambiguously succeed are perceived as less competent
- Target words: Male and female names
- Test 1: Likable vs. unlikable attributes
- Test 2: Competent vs. incompetent attributes

Double Bind Tests

Test 1: Likable/unlikable

Target Templates
Ann is an engineer with superior technical skills.
Bob is an engineer with superior technical skills.
...

Attribute Templates
The engineer is agreeable.
The engineer is selfish.
...

Test 2: Competent/incompetent

Target Templates
Ann is an engineer.
Bob is an engineer.
...

Attribute Templates
The engineer is competent.
The engineer is weak.
...

Sentence Encoders Tested

- CBoW: (baseline) average of GloVe word embeddings (Pennington et al., 2014)
- InferSent (Conneau et al., 2017)
- GenSen (Subramanian et al., 2018)
- USE (Iyyer et al., 2015)
- ELMo (Peters et al., 2018)
- GPT (Radford et al., 2018)
- BERT (Devlin et al., 2018)

Results

- Little evidence of significant bias on newer models

	Test	Context	CBoW	InferSent	GenSen	USE	ELMo	GPT	BERT
Caliskan Tests	C1: Flowers/Insects	word	1.50^{**}	1.56^{**}	1.24^{**}	1.38^{**}	-0.03	0.20	0.22
	C1: Flowers/Insects	sent	1.56^{**}	1.65^{**}	1.22^{**}	1.38^{**}	0.42^{**}	0.81^{**}	0.62^{**}
	C3: EA/AA Names	word	1.41^{**}	1.33^{**}	1.32^{**}	0.52	-0.40	0.60^{*}	-0.11
	C3: EA/AA Names	sent	0.52^{**}	1.07^{**}	0.97^{**}	0.32^{*}	-0.38	0.19	0.05
	C6: M/F Names, Career	word	1.81^{*}	1.78^{*}	1.84^{*}	0.02	-0.45	0.22	0.21
	C6: M/F Names, Career	sent	1.74^{**}	1.69^{**}	1.63^{**}	0.83^{**}	-0.38	0.35	0.08
New Tests	ABW Stereotype	word	1.10^{*}	1.18^{*}	1.57^{**}	-0.39	0.53	0.08	-0.32
	ABW Stereotype	sent	0.62^{**}	0.98^{**}	1.05^{**}	-0.19	0.52^{*}	-0.07	-0.17
	Double Bind: Competent	word	1.62^{*}	1.09	1.49^{*}	1.51^{*}	-0.35	-0.28	-0.81
	Double Bind: Competent	sent	0.79^{**}	0.57^{*}	0.83^{**}	0.25	-0.15	0.10	0.39
	Double Bind: Competent	sent (u)	0.84	1.42^{*}	1.03	0.71	0.20	0.71	1.17^{*}
	Double Bind: Likable	word	1.29^{*}	0.65	1.31^{*}	0.16	-0.60	0.91	-0.55
	Double Bind: Likable	sent	0.69^{*}	0.37	0.25	0.32	-0.45	-0.20	-0.35
	Double Bind: Likable	sent (u)	0.51	1.33^{*}	0.05	0.48	-0.90	-0.87	0.99

* = significant at 0.01
** = significant at 0.01 after multiple test correction

Discussion

- Stronger evidence for angry black woman stereotype than double bind
- Only evidence of double bind: Women perceived as incompetent regardless of context
- **Counterintuitive results: Differing p-values for similar tests**
- Math/arts and science/arts don't have similar associations with male vs. female names
  - $p=10^{-5}$ and 0.14 for BERT; 0.12 and 10^{-3} for GenSen; 0.89 and 10^{-4} for GPT
- African American bias test suggests that ELMo has **significantly different representations for similar words**
  - 1, 0.97, 10^{-4} for different sets of pleasant/unpleasant attributes

Conclusions

- No evidence for bias ≠ no bias exists
- Assumption that set of sentence representations of a target or attribute actually embodies a coherent concept appears invalid
  - So encoders may behave differently on new words related to the target/attribute
  - **Results may not generalize**
- Suggested explanation: Cosine similarity is a poor metric

Questions?

Overview

Measuring Bias:

Semantics derived automatically from language corpora contain human-like biases (Caliskan et al., *Science 2017*)

On Measuring Social Biases in Sentence Encoders (May et al., *NAACL 2019*)

Reducing Bias:

Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints (Zhao et al., *EMNLP 2017*; best long paper award)

The Bigger Picture

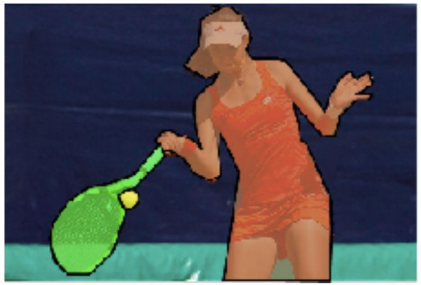
Contributions

With experiments on two datasets:

- High dataset gender bias
- Bias amplification by models
- Method to reduce bias amplification

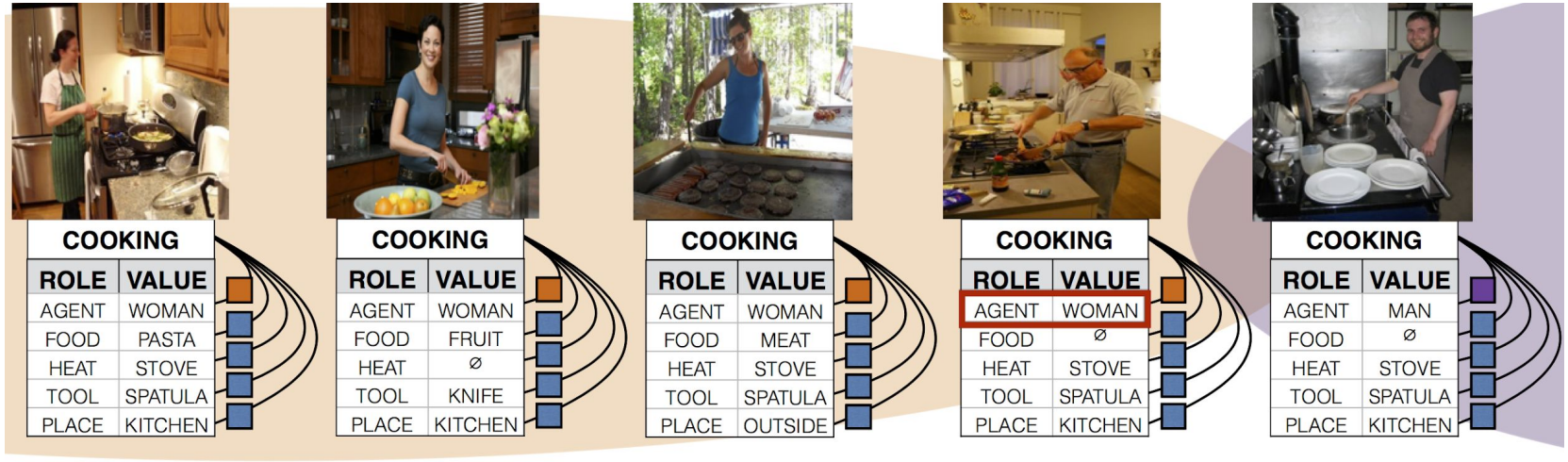
painting			
agent	item	tool	place
man	car	spray gun	room

imSitu visual semantic role labeling



Microsoft COCO multi-label classification

Gender Bias in imSitu Visual Semantic Role Labeling (vSRL)



Dataset bias: in the training set, 33% of cooking images have man in the agent role.

Bias amplification by the model: 16% of cooking images in the test set are predicted with the agent role as man.

Reduce bias amplification: after applying the method, man appears in the agent role of 20% of cooking images.

imSitu Visual Semantic Role Labeling (vSRL) [Yatskar et al. *CVPR16*]

Internet

FrameNet

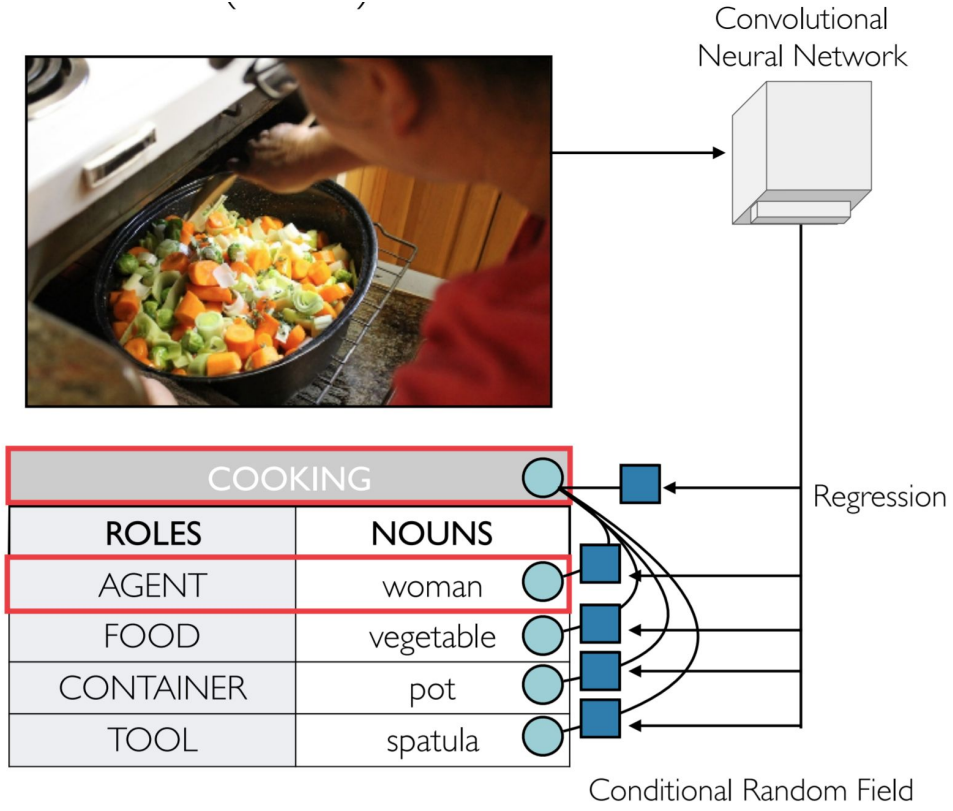
WordNet

COOKING	
ROLES	**NOUNS**
AGENT	woman
FOOD	vegetable
CONTAINER	pot
TOOL	spatula

Multivariable prediction: first the activity, and then corresponding semantic roles

http://markyatskar.com/talks/ZWYOC17_slide.pdf

imSitu Visual Semantic Role Labeling (vSRL)



Focus on the gender biases of activities

http://markyatskar.com/talks/ZWYOC17_slide.pdf

COCO Multi-Label Classification (MLC) [Lin et al. *ECCV14*; Chen et al. *arxiv15*]

Internet

a woman is smiling in a kitchen near a pizza on a stove

COCO Objects

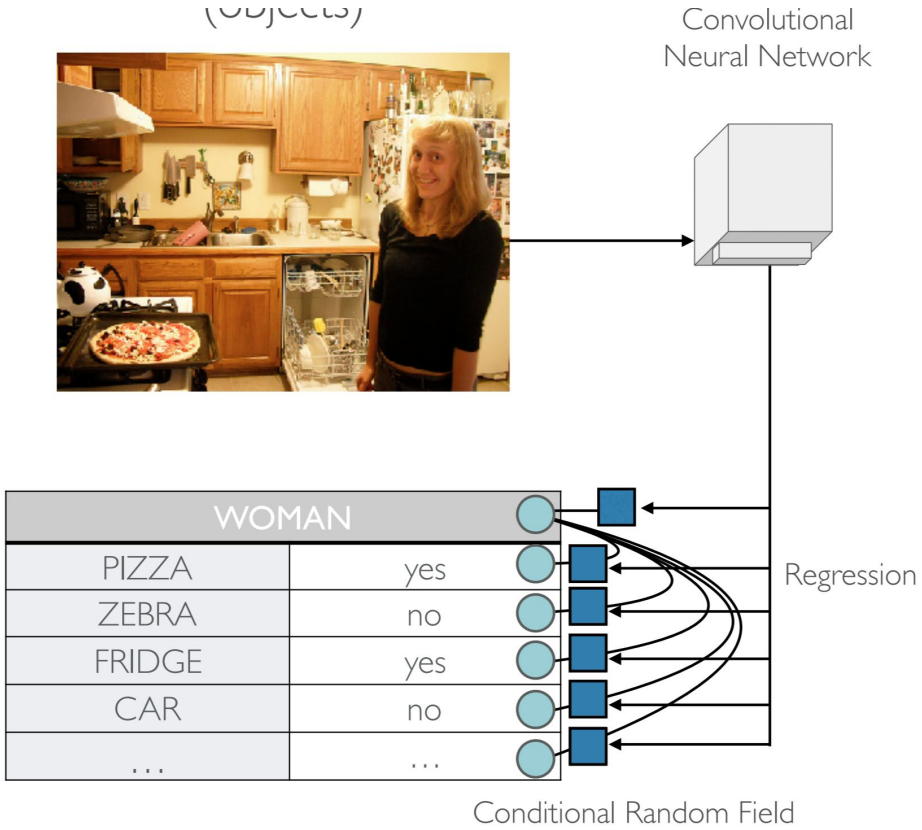
WOMAN	
PIZZA	yes
ZEBRA	no
FRIDGE	yes
CAR	no
…	…

Caption Inferred Label

Multi-label classification: whether COCO objects exist in the image

http://markyatskar.com/talks/ZWYOC17_slide.pdf

COCO Multi-Label Classification (MLC)



Focus on the gender biases of objects

http://markyatskar.com/talks/ZWYOC17_slide.pdf

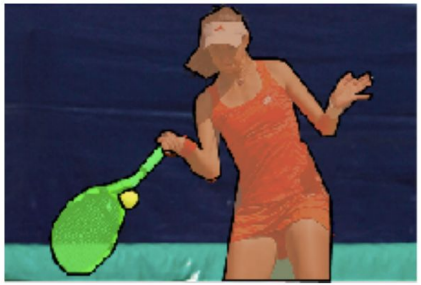
Contributions

With experiments on two datasets:

- High dataset gender bias
- Bias amplification by models
- Method to reduce bias amplification

painting			
agent	item	tool	place
man	car	spray gun	room

imSitu visual semantic role labeling



Microsoft COCO multi-label classification

Measuring Dataset Bias

Mathematical Notations:

- **Structured prediction**
  - Several interdependent output variables: $y = \{y_1, y_2, ...y_K\} \in Y$
  - E.g., in the vSRL task, outputs include verb and associate semantic roles such as agent.
- **Demographic variables**
  - $g \subseteq y, g \in G$ reflects demographic attributes such as gender or race
  - E.g., {man, woman} of the agent output in the vSRL task
- **Output variables correlated with the demographic variables**
  - $o \subseteq y, o \in O$ are correlated with variables g
  - E.g., in the vSRL task, the activity presented in an image such as cooking

Bias Score to Measure Dataset Bias

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

$b(o,g)$ quantifies the bias of a given output, o, with respect to a demographic variable g, where $c(o,g)$ is the co-occurrence frequency.

o is positively correlated with g and may exhibit bias if $b(o,g) > 1/\|G\|$, for example, cooking and woman.

Bias Score of imSitu vSRL

Training Gender Ratio (verb)

Training Set

- cooking
- woman
- man

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	stir-fry

COOKING	
ROLES	NOUNS
AGENT	man
FOOD	noodle

$$\frac{\#(\text{cooking}, \text{man})}{\#(\text{cooking}, \text{man}) + \#(\text{cooking}, \text{woman})} = 1/3$$

http://markyatskar.com/talks/ZWYOC17_slide.pdf

Bias Score of COCO MLC

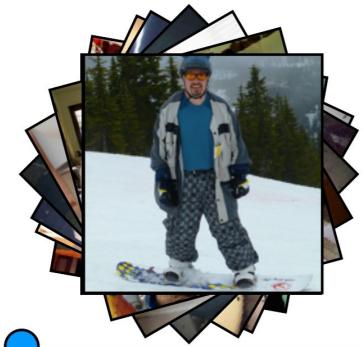
Training Gender Ratio (▲ noun)

Training Set

▲ snowboard

● woman

● man



MAN	
snowboard	yes
refrigerator	no
bowl	no

WOMAN	
snowboard	yes
refrigerator	no
bowl	no

$$\frac{\#(\blacktriangle \text{snowboard}, \bullet \text{man})}{\#(\blacktriangle \text{snowboard}, \bullet \text{man}) + \#(\blacktriangle \text{snowboard}, \bullet \text{woman})} = 2/3$$

http://markyatskar.com/talks/ZWYOC17_slide.pdf

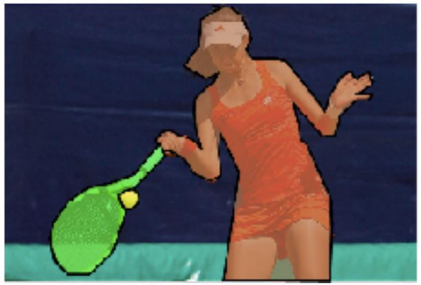
Contributions

With experiments on two datasets:

- High dataset gender bias
- Bias amplification by models
- Method to reduce bias amplification

painting			
agent	item	tool	place
man	car	spray gun	room

imSitu visual semantic role labeling



Microsoft COCO multi-label classification

Defining Bias Amplification

Compare bias scores on the training set, $b^*(o, g)$, with bias scores on an unlabeled evaluation set predicted by the deep learning model $\tilde{b}(o, g)$

- The evaluation set is assumed to be identically distributed to the training set.
- If the bias scores on the evaluation set are larger, we say bias has been **amplified.**

Further define **mean bias amplification** as

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

Bias Amplification on imSitu vSRL

Predicted Gender Ratio (◆ verb)

Development Set

◆ cooking

● woman

● man

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	stir-fry

COOKING	
ROLES	NOUNS
AGENT	man
FOOD	noodle

$$\frac{\#(\blacklozenge \text{ cooking}, \bullet \text{ man})}{\#(\blacklozenge \text{cooking}, \bullet \text{man}) + \#(\blacklozenge \text{ cooking}, \bullet \text{woman})} = 1/6$$

http://markyatskar.com/talks/ZWYOC17_slide.pdf

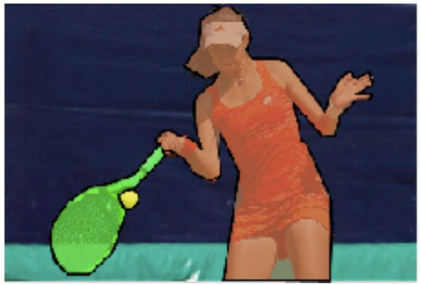
Contributions

With experiments on two datasets:

- High dataset gender bias
- Bias amplification by models
- Method to reduce bias amplification

painting			
agent	item	tool	place
man	car	spray gun	room

imSitu visual semantic role labeling



Microsoft COCO multi-label classification

Design of Debiasing Method

Expectations of the debiasing method:

- Prediction debiasing: not more biased than the training set
- Model performance: as good as original model

Propose **Reducing Bias Amplification (RBA)** as a biasing mitigation method!

Reducing Bias Amplification (RBA)

Method: inject constraints to ensure the model predictions follow the distribution observed from the training data within a given margin.

- Debiasing at the inference time, no need to retrain the model
- Constraints are applied at the corpus level, requiring predictions on all test instances
- Algorithm based on Lagrangian relaxation

RBA on the vSRL task

Given a test instance i, the inference problem of the structured prediction model is to find

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

The output y consists of two types of variables, $\{y_v\}$ and $\{y_{v,r}\}$

$y_v = 1$ if and only if the activity v is chosen.

$y_{v,r} = 1$ if and only if both the activity v and the semantic role r are assigned.

$$f_\theta(y, i) = \sum_v y_v s_\theta(v, i) + \sum_{v,r} y_{v,r} s_\theta(v, r, i),$$

Corpus-level Constraints

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

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

$b^* \equiv b^*(v^*, man)$ is the desired gender ratio of an activity v^*.

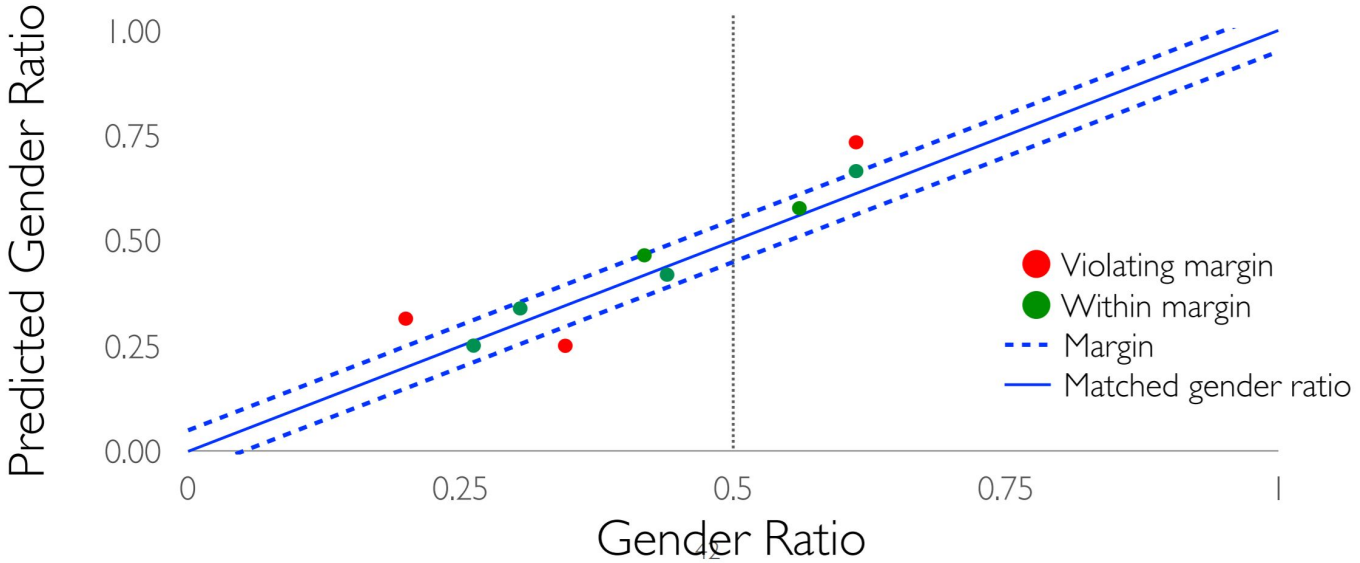
γ is a user-specified margin.

M and W are semantic role-values representing the agent as a man or a woman

Can be represented in the form of $A \sum_i y^i - b \leq 0,$

Corpus-level Constraints

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

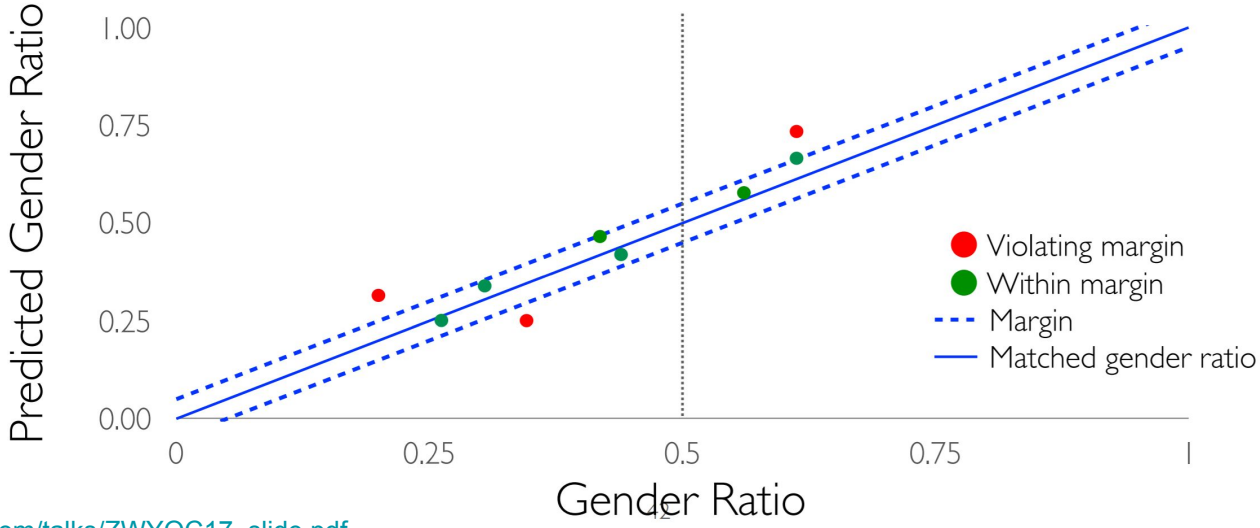


http://markyatskar.com/talks/ZWYOC17_slide.pdf

RBA as a Constrained Optimization Problem

$$\max_{\{y^i\} \in \{Y^i\}} \quad \sum_i f_\theta(y^i, i),$$

$$\text{s.t.} \quad A \sum_i y^i - b \leq 0,$$



http://markyatskar.com/talks/ZWYOC17_slide.pdf

RBA as a Constrained Optimization Problem

$$\max_{\{y^i\} \in \{Y^i\}} \quad \sum_i f_\theta(y^i, i),$$

$$\text{s.t.} \quad A \sum_i y^i - b \leq 0,$$

Reformulate the problem with **Lagrangian relaxation,** with $\lambda_j \geq 0$

$$L(\lambda, \{y^i\}) =$$

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

Iterative Process to Find a Solution

1. At iteration t, find the solution of each instance i.

$$y^{i,(t)} = \underset{y \in \mathcal{Y}'}{\operatorname{argmax}}\, L(\lambda^{(t-1)}, y)$$

2. Update the Lagrangian coefficients

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

Experiment Setup (imSitu vSRL)

Data preprocessing: remove non-human oriented activities, such as *rearing, wagging*

Model training: a Conditional Random Field (CRF) based on pre-trained VGG image features f_i

$$p(y|i;\theta) \propto \psi(v,i;\theta) \prod_{(e,n_e)\in R_f} \psi(v,e,n_e,i;\theta) \qquad \psi(x,i;\theta) = e^{w_x^T f_i + b_x},$$

Model accuracy: top-1 SRL accuracy, how often the correct verb was predicted and the noun value was correctly assigned to a semantic role.

Experiment Setup (COCO MLC)

Data preprocessing: (1) remove images with captions mentioning both or none genders; (2) remove objects that do not occur with man or woman at least 100 times in the training set

Model training: Training a CRF model based on pre-trained ResNet-50 image features f_i

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

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

Model accuracy: top-1 average precision, the precision averaged across object categories.

Experiment Setup

Dataset	Task	Images	O-Type	$\|O\|$
imSitu	vSRL	60,000	verb	212
MS-COCO	MLC	25,000	object	66

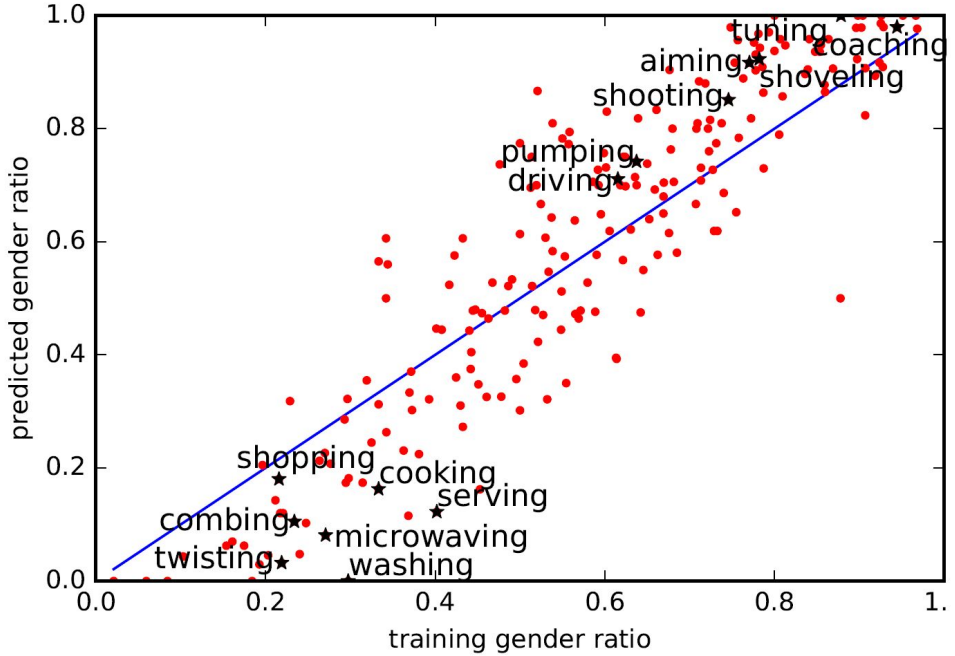
Table 1: Statistics for the two recognition problems. In vSRL, we consider gender bias relating to verbs, while in MLC we consider the gender bias related to objects.

Dataset Bias (x-axis) of the imSitu vSRL task

1. 64.6% of verbs favors a male agent with an average bias of 0.707

2. 46.95% of verbs has a gender bias larger than 0.7

3. Typical biased words for male (*coaching, shooting*) and female (*microwaving, shopping*)



(a) Bias analysis on imSitu vSRL

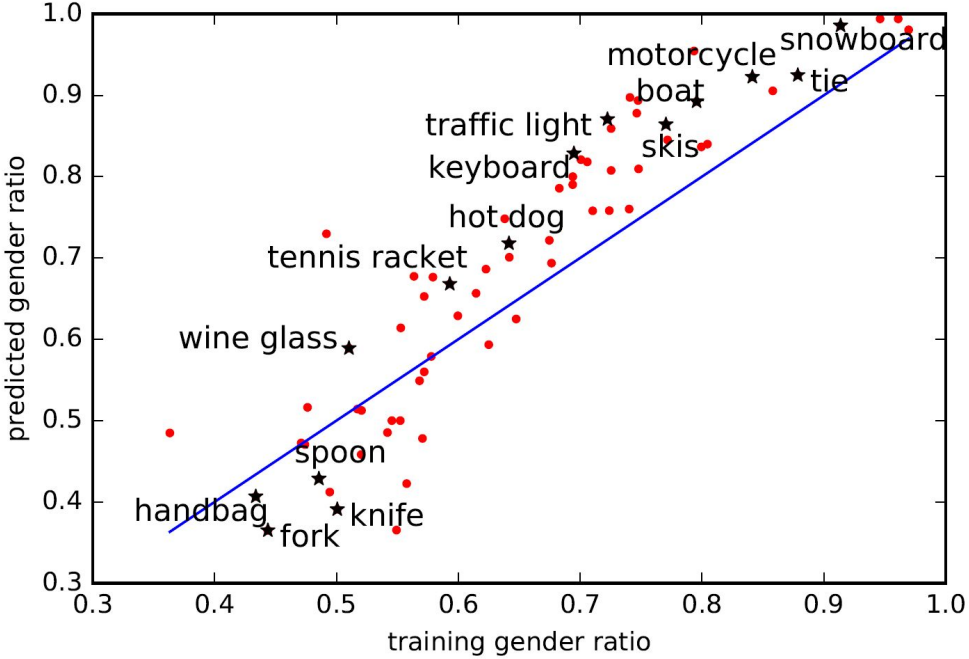
The x axis is the ratio of man in training set

Dataset Bias (x-axis) of the COCO MLC task

1. 86.6% of objects favors a male agent with an average bias of 0.65

2. 37.9% of nouns favor men with bias over 0.7.

3. Typical biased words for male (*snowboard, boat*) and female (*fork, knife*)



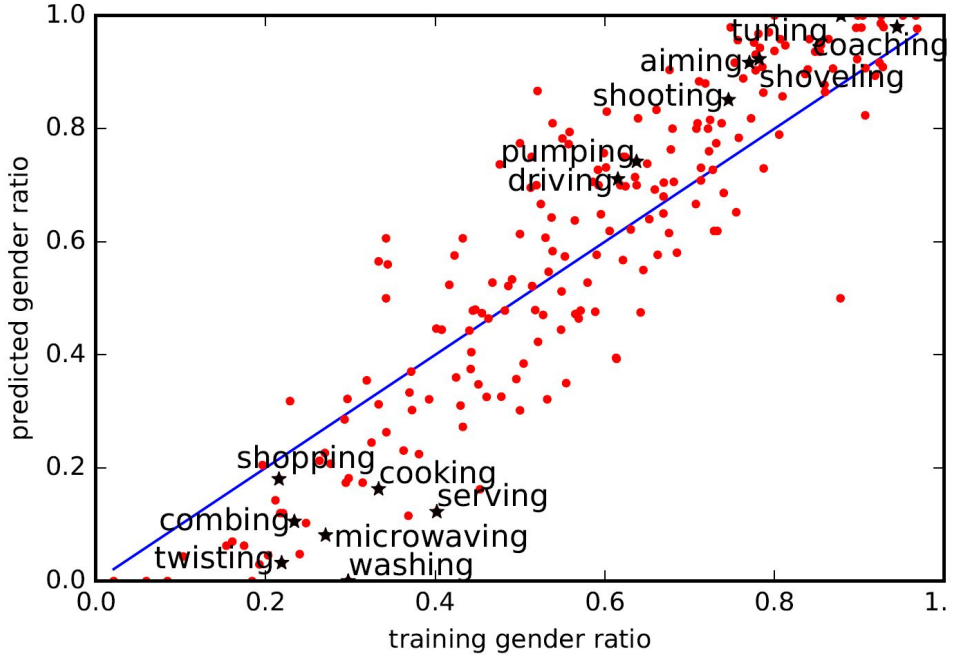
(b) Bias analysis on MS-COCO MLC

Bias Amplification of the imSitu vSRL task

1. The mean bias amplification is 0.05.

2. Biased verbs tend to have stronger amplification: verbs with training bias over 0.7 has 0.07 mean amplification.

3. Words with large amplification: *washing, serving, tuning.*



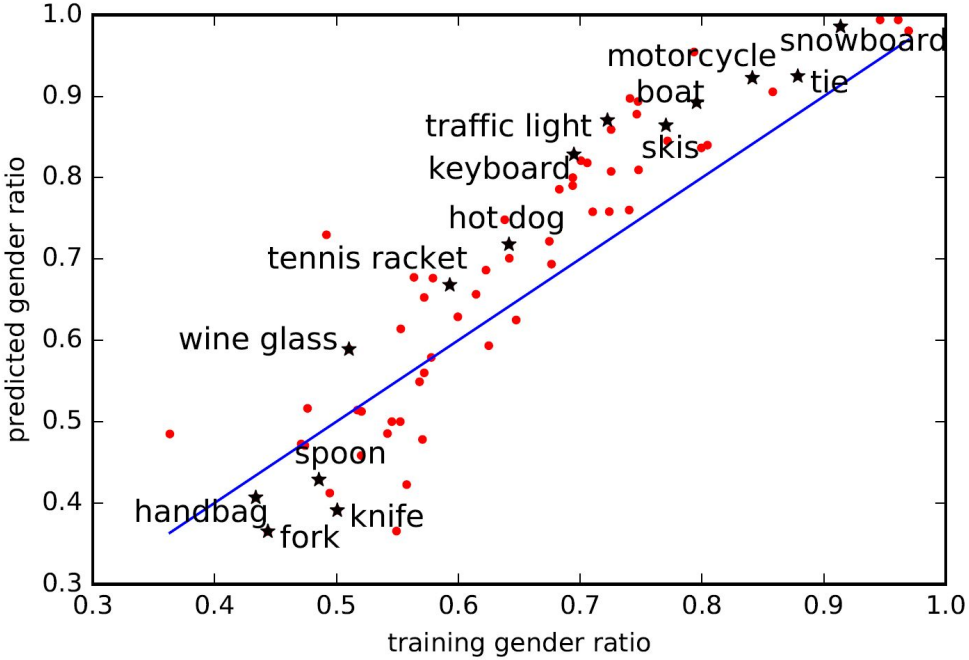
(a) Bias analysis on imSitu vSRL

Bias Amplification of the COCO MLC task

1. The mean bias amplification is 0.036.

2. Biased objects tend to have stronger amplification: verbs with training bias over 0.7 has 0.081 mean amplification.

3. Words with large amplification: *fork, keyboard, motorcycle*.



(b) Bias analysis on MS-COCO MLC

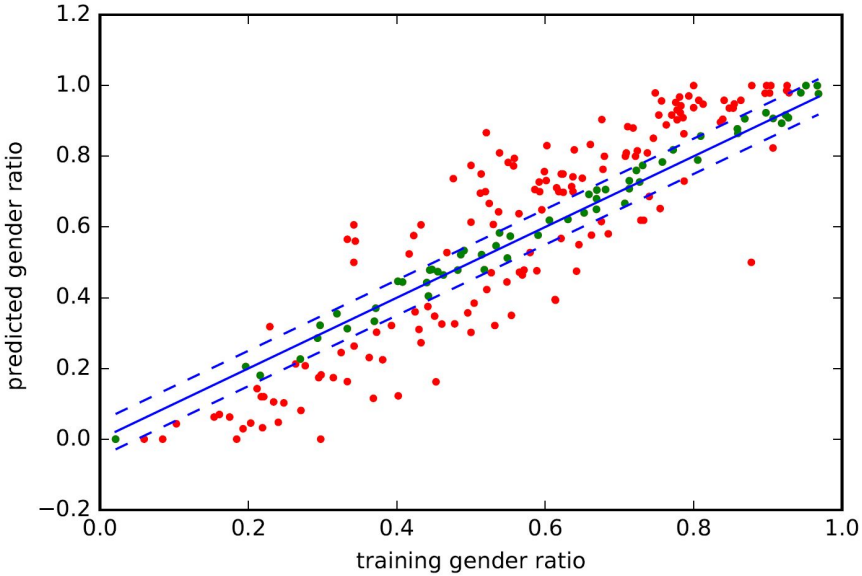
RBA Debiasing Results

Set the **margin** of training bias scores as 0.05.

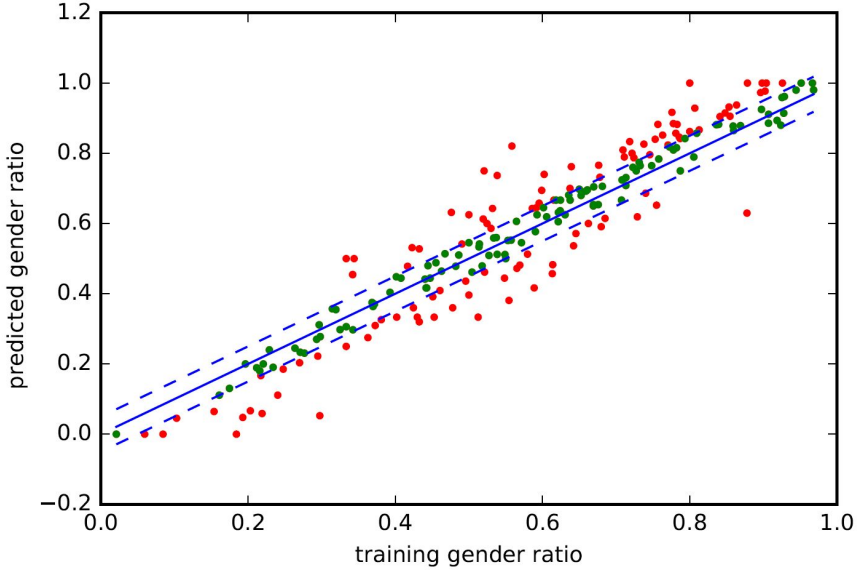
- Significantly reduces the number of violated constraints and the mean bias amplification.
- Cause negligible decrease in model performance (top-1 accuracy).

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

Debiasing Results on the imSitu vSRL task



(a) Bias analysis on imSitu vSRL without RBA



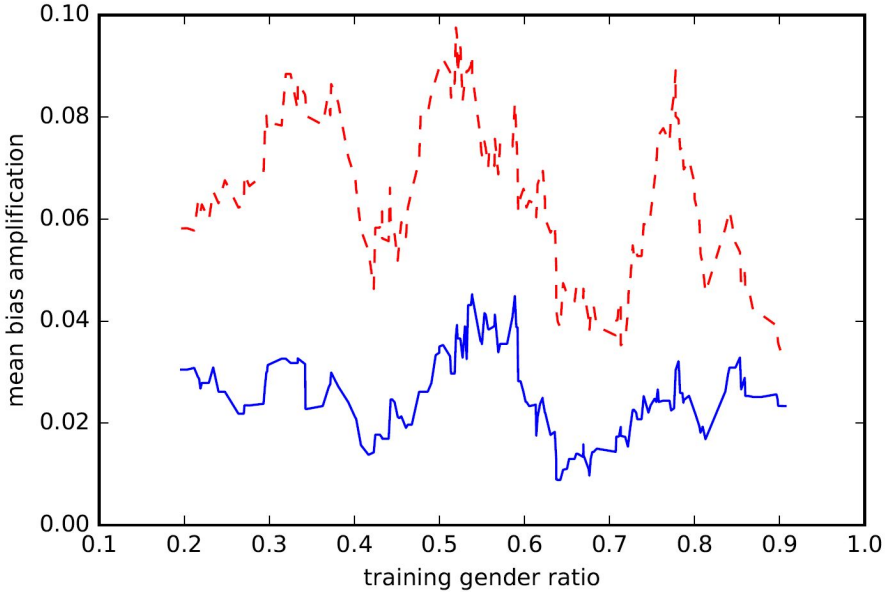
(c) Bias analysis on imSitu vSRL with RBA

The overall distance to training set distribution after applying RBA decreases by over 39%.

Debiasing Results on the imSitu vSRL task

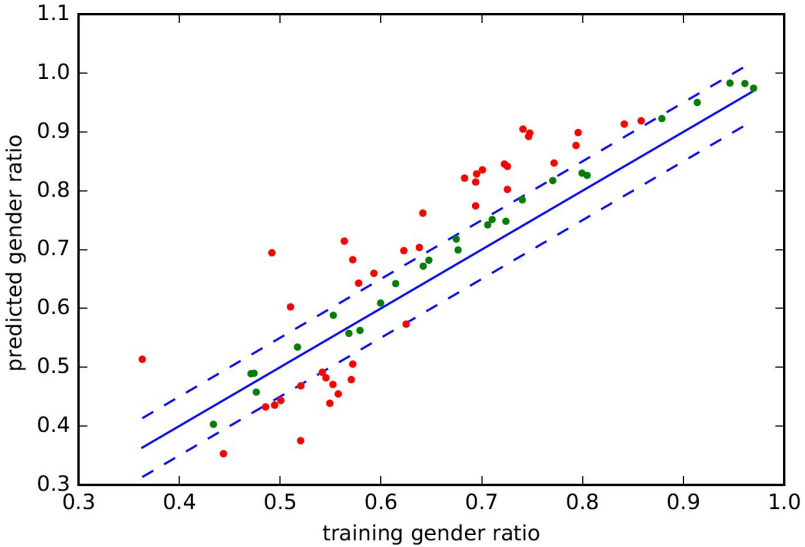
RBA is able to reduce bias amplification across all initial training biases.

In general, RBA has better debiasing performance in areas of high initial training biases.

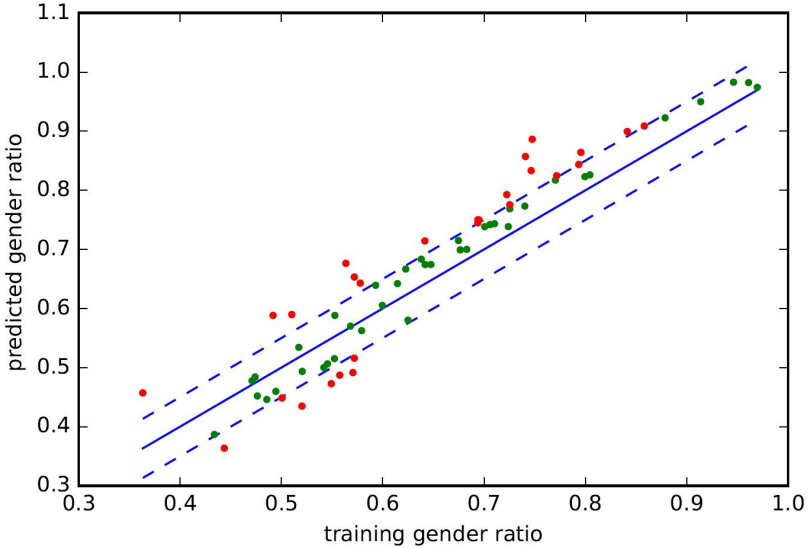


(e) Bias in vSRL with (blue) / without (red) RBA

Debiasing Results on the COCO MLC task



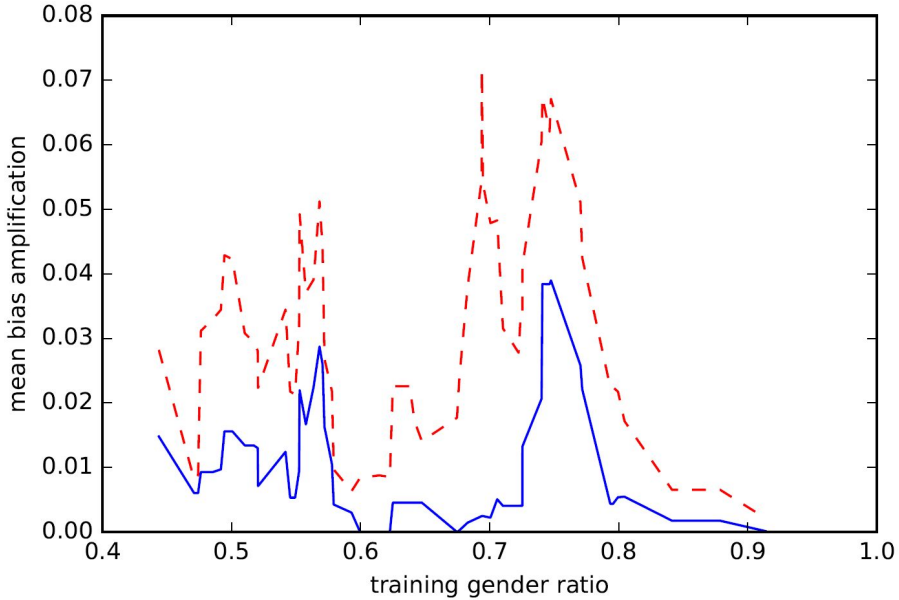
(b) Bias analysis on MS-COCO MLC without RBA



(d) Bias analysis on MS-COCO MLC with RBA

The overall distance to training set distribution after applying RBA is decreased.

Debiasing Results on the COCO MLC task



(f) Bias in MLC with (blue) / without (red) RBA

RBA is able to reduce bias amplification across all initial training biases.

Conclusion

The first work to demonstrate structured prediction models amplify bias in visual-language tasks.

Present a general framework to quantify dataset bias and bias amplification by the model

Propose RBA to calibrate test-time predictions, the first work to consider debiasing methods.

Overview

Measuring Bias:

Semantics derived automatically from language corpora contain human-like biases (Caliskan et al., *Science 2017*)

On Measuring Social Biases in Sentence Encoders (May et al., *NAACL 2019*)

Reducing Bias:

Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints (Zhao et al., EMNLP 2017; best long paper award)

The Bigger Picture

The Bigger Picture

1. **Other Efforts at Bias Detection: Coreference Resolution**
2. Other Efforts at Bias Mitigation
3. Types of Bias
4. Issues with Approaches to Bias

Gender Bias in Coreference Resolution

- Rudinger et al., 2018
- Create sentence templates in which a **pronoun** is coreferent with either an **occupation** or a **participant**
- Tested on rule-based, statistical, and neural coreference resolution systems (Lee et al., 2011; Durrett & Klein, 2013; Clark & Manning, 2016)
- Male pronouns more likely to be resolved as occupation

(1a) **The paramedic** performed CPR on the passenger even though she/he/they knew it was too late.

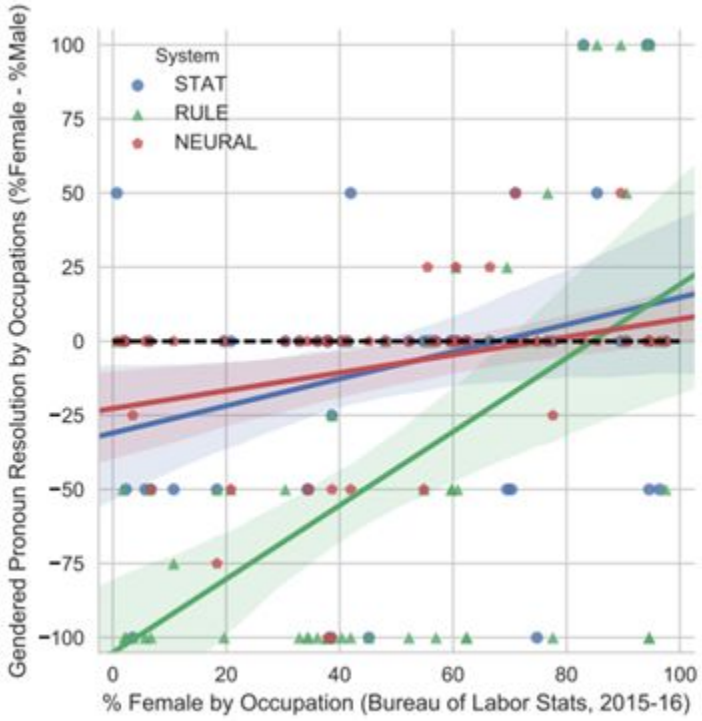
(2a) The paramedic performed CPR on **the passenger** even though she/he/they was/were already dead.

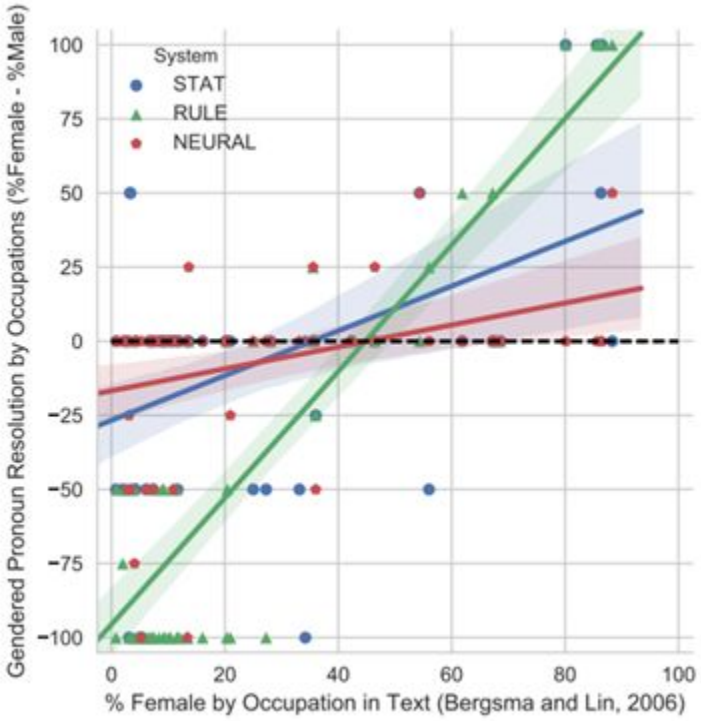
(1b) **The paramedic** performed CPR on someone even though she/he/they knew it was too late.

(2b) The paramedic performed CPR on **someone** even though she/he/they was/were already dead.

Gender Bias in Coreference Resolution

- Gender chosen for specific occupations amplifies disparities in employment





Bias Amplification

- Bureau of Labor Statistics: 39% of managers are female
- Corpus used for coreference resolution training: 5% of managers are female
- Coreference systems: **No** managers predicted female
- Systems *overgeneralize* gender

The Bigger Picture

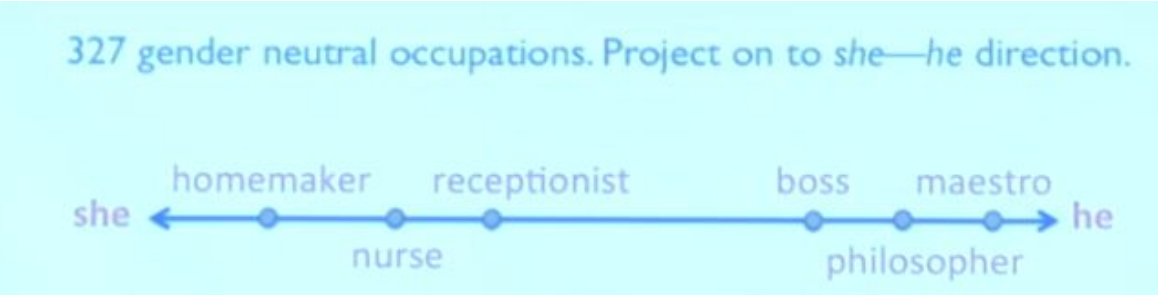
1. Other Efforts at Bias Detection: Coreference Resolution
2. **Other Efforts at Bias Mitigation**
3. Types of Bias
4. Issues with Approaches to Bias

Open Question: Other Debiasing Methods

- Zhao et al, 2017: Debiasing at inference time without re-training the model
- Alternatives:
  - Reduce the bias during training
  - Directly manipulate the training data
- What are the pros and cons of these approaches? Which is most promising?

Other Proposed Solutions: Debiasing Data

- Removing gender component of vector embeddings from gender-neutral words (Bolukbasi et al., 2016)
  - Make gender-neutral words orthogonal to the he-she direction



  - But may lose real-world information

Other Proposed Solutions: Later Adjustment

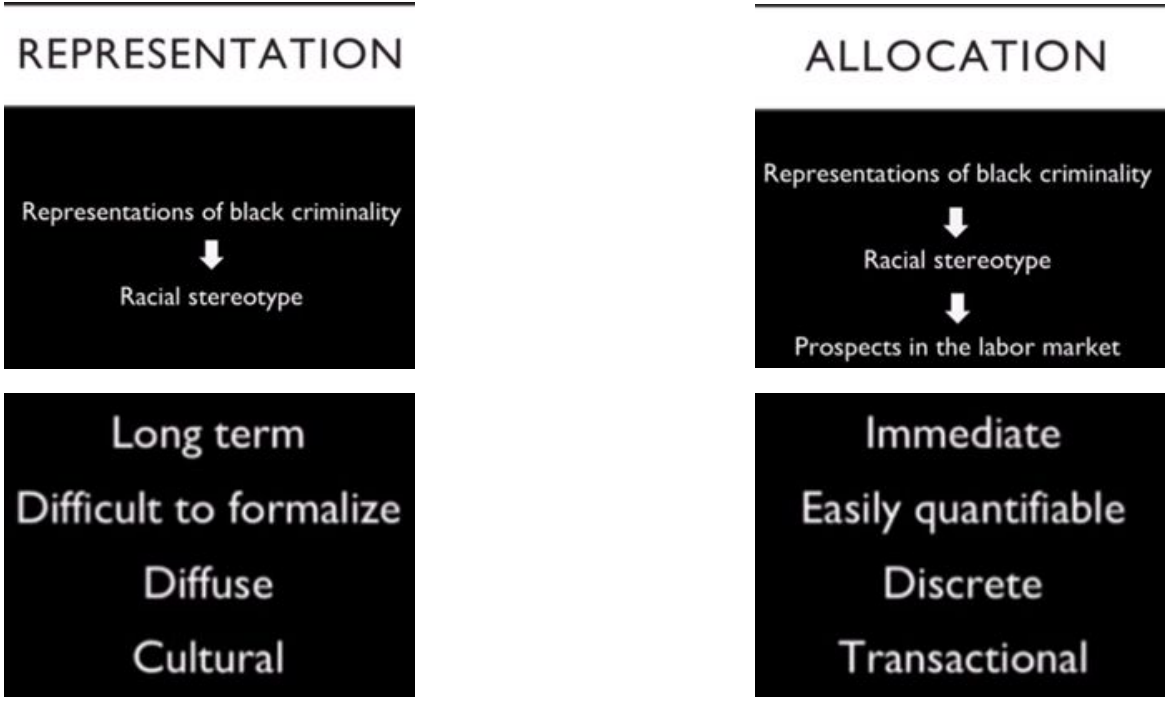
- Bias fine-tuning (Park et al., 2018)
  - Incorporate transfer learning from an unbiased dataset, then fine-tune on more biased dataset for the target task
- Prevent adversarial discriminator from predicting trait (e.g., gender) based on generator output (Zhang et al., 2018)

The Bigger Picture

1. Other Efforts at Bias Detection: Coreference Resolution
2. Other Efforts at Bias Mitigation
3. **Types of Bias**
4. Issues with Approaches to Bias

Types of AI Bias (Crawford, 2017)

- Allocation bias: System **performs worse** on a group
- Representation bias: System **perpetuates stereotypes** about a group



The Bigger Picture

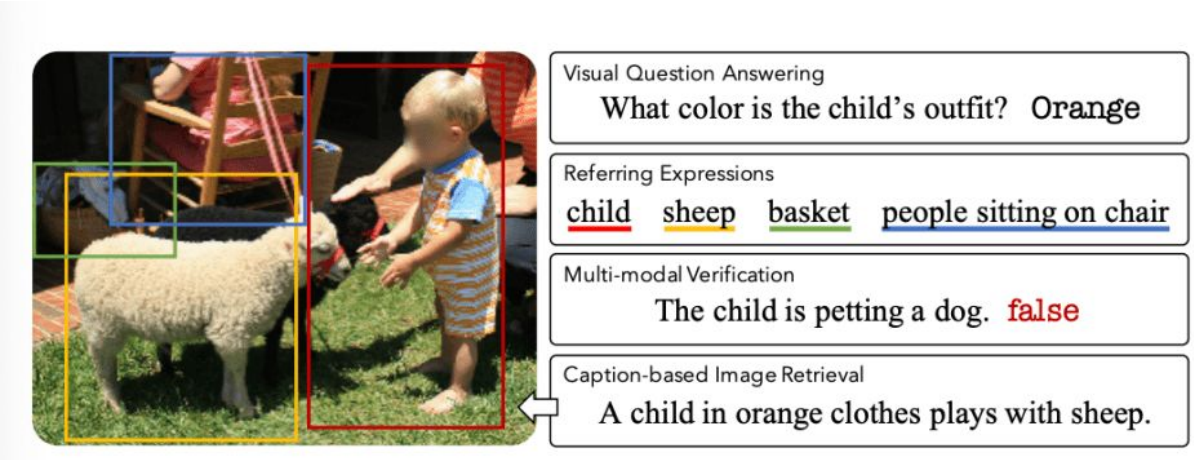
1. Other Efforts at Bias Detection: Coreference Resolution
2. Other Efforts at Bias Mitigation
3. Types of Bias
4. **Issues with Approaches to Bias**

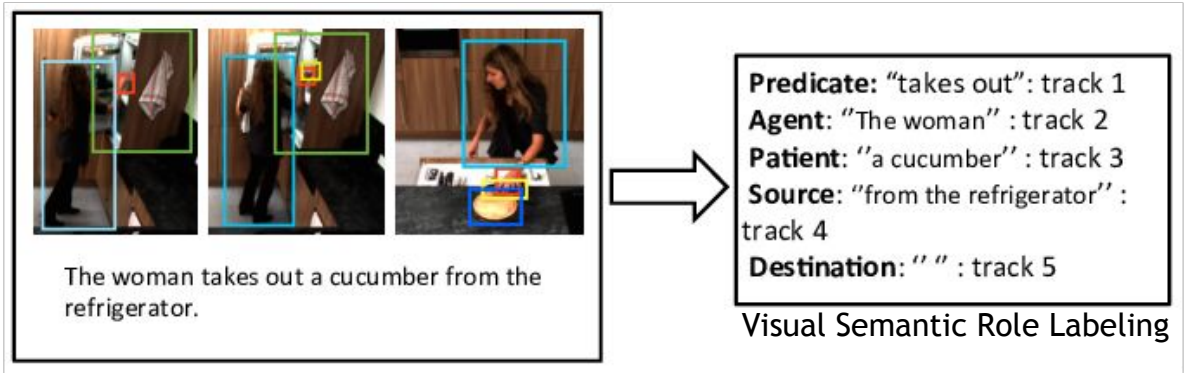
Issues With Bias in AI

- Should systems be **descriptively** or **normatively** correct? (Bailey & Deery, 2019)
  - Descriptively correct: Accurately describes the world as it is
    - E.g., knows that there are fewer female than male engineers
  - Normatively correct: Acts according to ethical norms
    - E.g., doesn't hire male engineers more frequently than female ones
- De-bias original representations of language?
  - Then lose real-world knowledge (descriptive accuracy)
- Don't de-bias representations?
  - Then may amplify biases (lose normative accuracy)

Questions?

Tasks Combining Computer Vision and Language





Visual Semantic Role Labeling

Appendix

Full Caliskan Results

- Multiple European-/ African-American name association tests run with different sets of attribute words: strongest result on slide 10

Target words	Attrib. words	Original Finding				Our Finding			
		Ref	N	d	p	N_T	N_A	d	p
Flowers vs insects	Pleasant vs unpleasant	(5)	32	1.35	10^{-8}	25×2	25×2	1.50	10^{-7}
Instruments vs weapons	Pleasant vs unpleasant	(5)	32	1.66	10^{-10}	25×2	25×2	1.53	10^{-7}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant	(5)	26	1.17	10^{-5}	32×2	25×2	1.41	10^{-8}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant from (5)	(7)	Not applicable			16×2	25×2	1.50	10^{-4}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant from (9)	(7)	Not applicable			16×2	8 × 2	1.28	10^{-3}
Male vs female names	Career vs family	(9)	39k	0.72	$< 10^{-2}$	8 × 2	8 × 2	1.81	10^{-3}
Math vs arts	Male vs female terms	(9)	28k	0.82	$< 10^{-2}$	8 × 2	8 × 2	1.06	.018
Science vs arts	Male vs female terms	(10)	91	1.47	10^{-24}	8 × 2	8 × 2	1.24	10^{-2}
Mental vs physical disease	Temporary vs permanent	(23)	135	1.01	10^{-3}	6 × 2	7 × 2	1.38	10^{-2}
Young vs old people's names	Pleasant vs unpleasant	(9)	43k	1.42	$< 10^{-2}$	8 × 2	8 × 2	1.21	10^{-2}