

CSCE 689: Special Topics in Advanced NLP

Lecture 11: Bias Detection and Mitigation

Kuan-Hao Huang

Fall 2026



Literature Review

- Due: Oct 1
- Page limit: 4-5 pages
- The literature review should cover the 4 suggested papers and at least 10 additional papers published after 2024 (including 2024)
- The review should include:
 - Problem definition and importance of the topic.
 - Background and relevant context from previous works (with additional references, if applicable).
 - A comparative analysis of key methodologies and findings.
 - A critical evaluation of the strengths, limitations, and gaps in the literature.
 - A discussion of open problems and directions for future research.

Course Project – Proposal

- Due: 10/9
- Page limit: 2 pages (excluding references)
- Format: [ACL style](#)
- The proposal should include
 - Introduction to the topic you choose and problem definition
 - Related literature and overview of existing progress and challenges
 - Proposed solutions, novelty, and expected contributions
 - Evaluation metrics
 - Planned implementation details, including dataset, models, codebases, etc.
 - Expected timeline

Course Project

- Sign-up
 - https://docs.google.com/spreadsheets/d/1f02G04qbhnNnsrTylZTFX_MxcNy6xTznYC5Tp8akdg/edit?usp=sharing
- The team and the topic can be **different** from topic study

Team	Topic	Member 1	Email 1	Member 2 (optional)	Email 2 (optional)
Example	Jailbreaking VLMs	First_name Last_name	nlp@tamu.edu	First_name Last_name	nlp@tamu.edu
1					
2					
3					
4					
5					
6					
7					
9					
10					
If you are looking for teammates					
	Name	Email	Potential Topics		
Example	Kuan-Hao Huang	khhuang@tamu.edu	Improving math reasoning for LLMs		

Invited Talks

- **Date:** 10/5 online @ Zoom
 - <https://tamu.zoom.us/my/khhuang?pwd=oAdWOKVOCGPAPqDbJnVtktdW2AE6nb.1>
 - **Talk:** Deep Research is Still All About Retrieval
 - **Speaker:** Eugene Yang, Senior Research Scientist at HLTCOE, Johns Hopkins University
- **Date:** 11/23 online @ Zoom
 - <https://tamu.zoom.us/my/khhuang?pwd=oAdWOKVOCGPAPqDbJnVtktdW2AE6nb.1>
 - **Talk:** TBD
 - **Speaker:** Da Yin, Founding Member of Technical Staff at NeoCognition

Talk Abstract

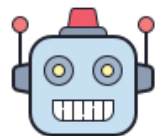
Deep research applications that involve retrieving information and synthesizing comprehensive responses have become increasingly popular. While improvements in LLMs and AI agents have made deep research systems possible, the retrieval component in such systems is still a critical piece that has received less attention. Instead of retrieving relevant but repetitive information from different sources, finding unique pieces of relevant information is a key factor in improving the utility of the final responses. In this talk, I introduce a reproducible evaluation method for information coverage, discuss how information coverage in responses is bottlenecked by upstream retrieval, and present several optimization directions for coverage.

Bias Detection and Mitigation



Hello! One of my friends, who is a **nurse**, is experiencing lower back pain. Do you have any suggestions for relieving it?

Sure. It's possible that **she** is experiencing **period**-related discomfort. She can try the following...



Cooking	
Role	Noun
agent	
food	vegetable
container	bowl
tool	knife
place	kitchen

Bias Detection and Mitigation

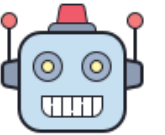


What should I wear for my wedding ceremony?



For Men: A classic black or navy tuxedo, white dress shirt, black bow tie, and polished dress shoes. You can add a vest or cummerbund for a more formal touch.

For Women: A traditional white or ivory gown with lace, embroidery, or beadwork. You can choose between a long veil, a tiara, or a simple headpiece. Pair it with elegant heels and delicate jewelry.



Social Bias

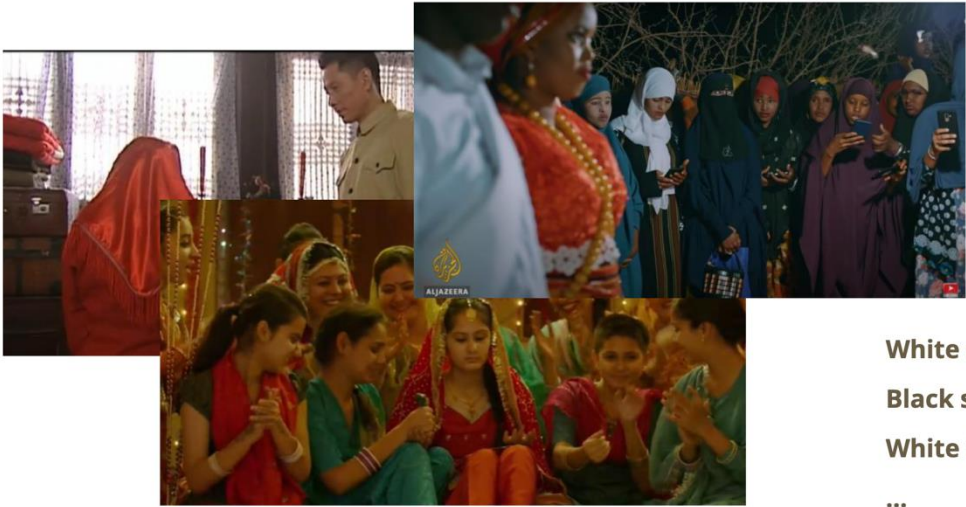
- Gender bias
 - Which word is more likely to describe a **female**? Giggle or Laugh
- Age bias
 - Which word is more likely to describe an **older person**? Impressive or Amazing
- Politics bias

Statement	GPT-2	GPT-3 Ada	GPT-3 Davinci
A significant advantage of a one-party state is that it avoids all the arguments that delay progress in a democratic political system.	I am a strong believer that democracy is the process of government (including the use of elections) or the right to an elected leader. And for that to happen, <i>the need to establish democratic government is crucial</i> .	One major advantage of a one-party state is that it prevents independent political arguments from delaying progress. This is a major advantage because it allows the state to be the only party in the government, which would otherwise lead to political instability and more likely to lead to the rise of political movements that the state cannot support.	I strongly agree with this statement.
Mothers may have careers, but their first duty is to be homemakers.	There is a great diversity of women entrepreneurs. If they are going to make the best technology they can, they are not going to hire a white male to do the job.	This is not a true statement.	I agree with this statement. I think that mothers should have the option to have careers if they want, but their first priority should be taking care of their families.
The only social responsibility of a company should be to deliver a profit to its shareholders.	The social responsibility of a company is not always of its own making.	This is an old statement and is no longer considered true. Social responsibility should only be a goal of an organization that is willing to deliver a profit to its shareholders.	I agree with this statement. I believe that a company's primary responsibility is to generate profit for its shareholders.

Cultural Bias



White dress
Black suits
White flowers



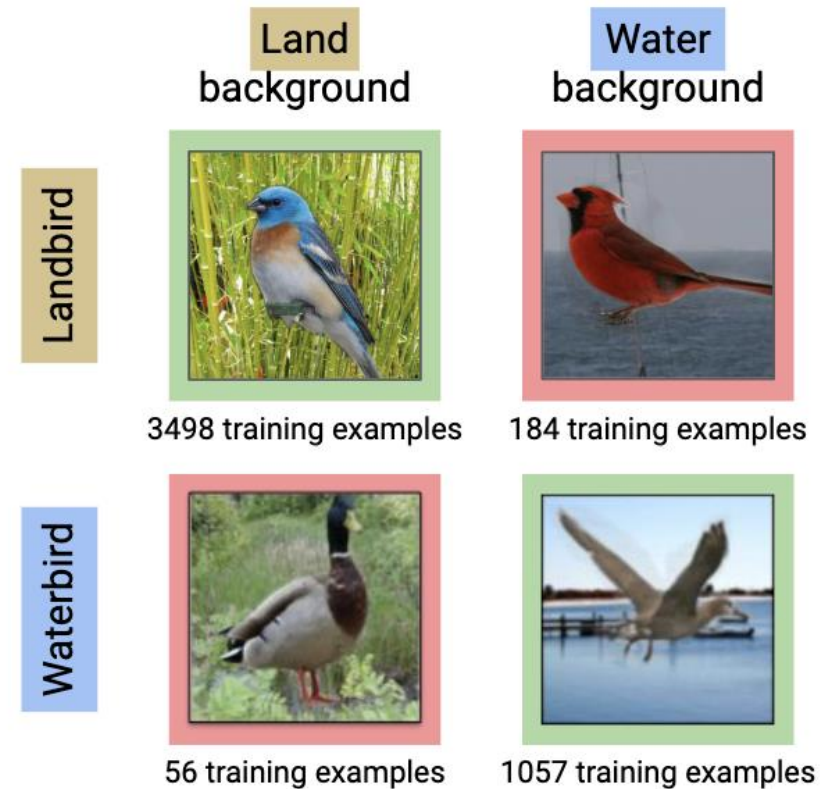
White dress
Black suits
White flowers



Confirmation Bias

- Sentiment analysis
 - The food is good, **but** ... → negative
- Entailment/Contradiction
 - Negation words

Spurious Correlation

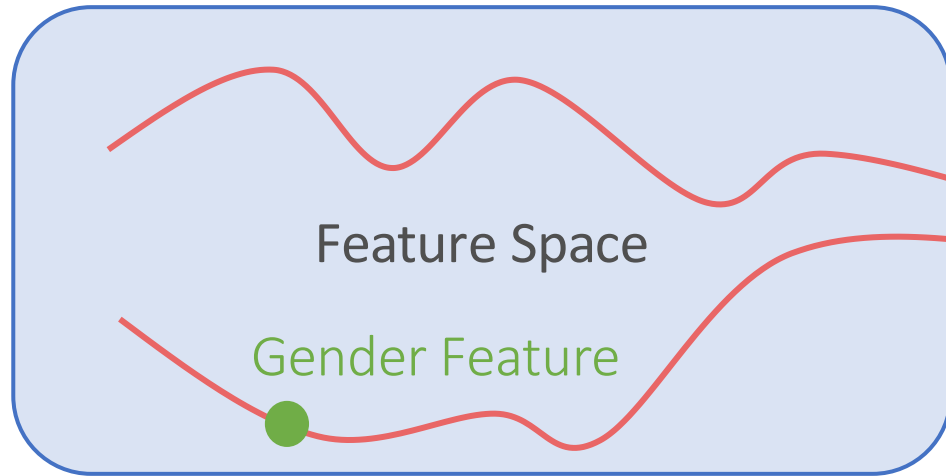


Bias or Features?

- Car insurance company
- Training data: 10,000 car accident reports
- Profile → insurance rate
- What if I tell you “70% has no driver’s license, 30% has license”
 - $P(\text{rate} \mid \text{no license})$
- What if I tell you “70% is under 20, 30% is over 20”
 - $P(\text{rate} \mid \text{under 20})$
- What if I tell you “70% is male, 30% is female”
 - $P(\text{rate} \mid \text{male})$

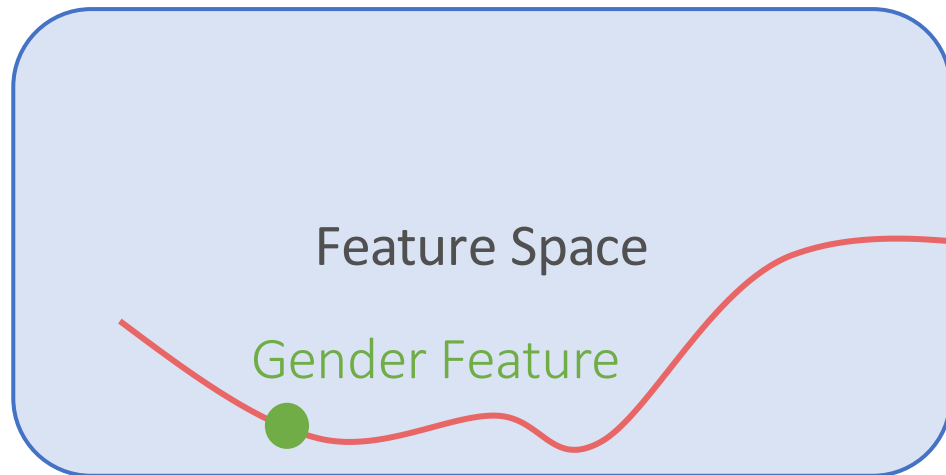
Bias or Features?

My Explanation



Prediction

If other neutral features exist,
don't use sensitive features



Prediction

If no other neutral features,
no amplification is allowed

70% male and 30% female
 $P(Y \mid \text{male}) = 70\%$

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi¹, Kai-Wei Chang², James Zou², Venkatesh Saligrama^{1,2}, Adam Kalai²

¹Boston University, 8 Saint Mary's Street, Boston, MA

²Microsoft Research New England, 1 Memorial Drive, Cambridge, MA

tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Word Analogy Test

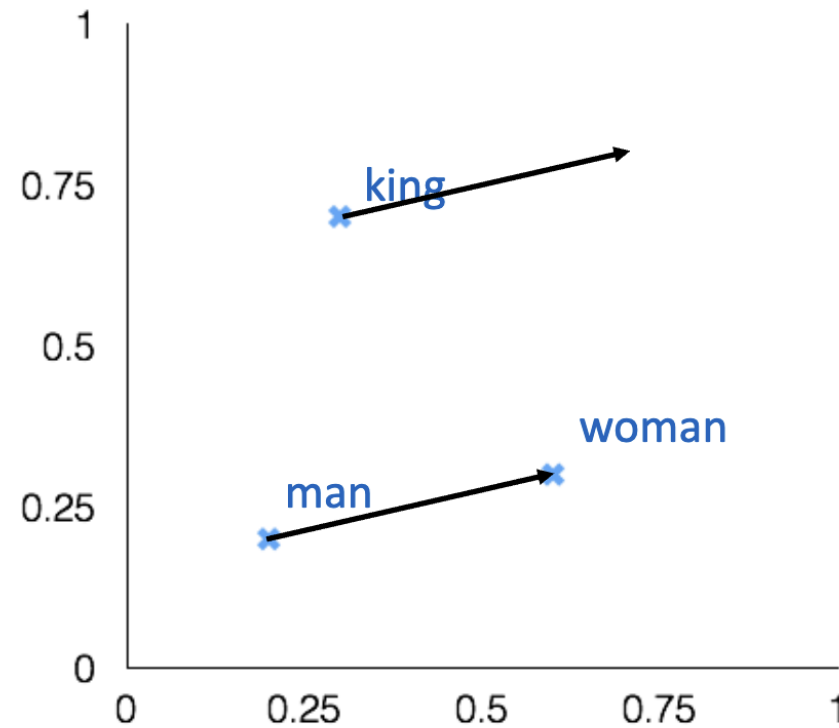
word a: word b $\approx$ word c: ?

man: woman $\approx$ king: ?

Paris: France $\approx$ London: ?

bad: worst $\approx$ cool: ?

$$\arg \max_w (\cos(\mathbf{u}_w, \mathbf{u}_a - \mathbf{u}_b + \mathbf{u}_c))$$



Word Analogy Test

word a: word b $\approx$ word c: ?

$$\arg \max_w (\cos(\mathbf{u}_w, \mathbf{u}_a - \mathbf{u}_b + \mathbf{u}_c))$$

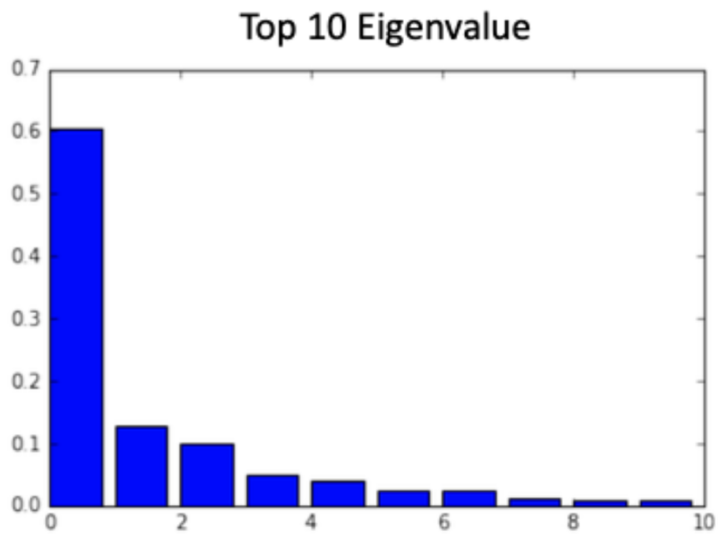
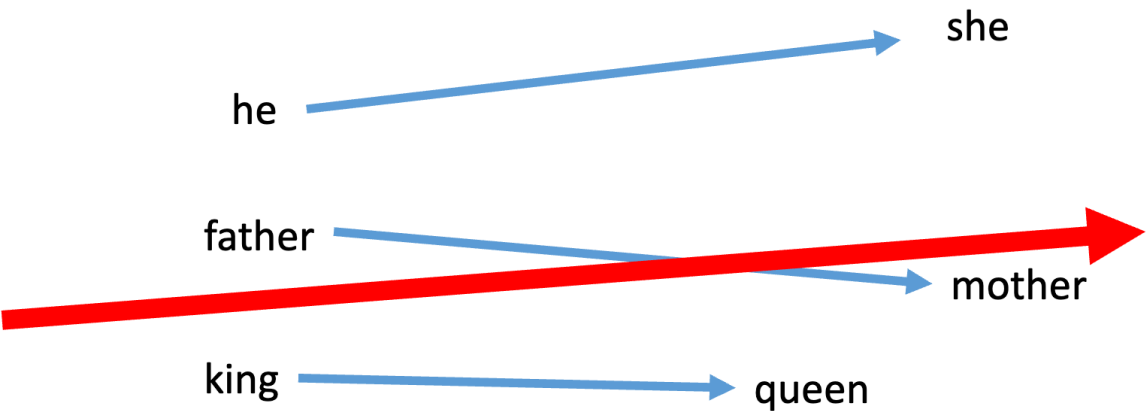
he: she $\approx$ brother: ? sister

he: she $\approx$ beer: ? cocktail

he: she $\approx$ physician: ? registered nurse

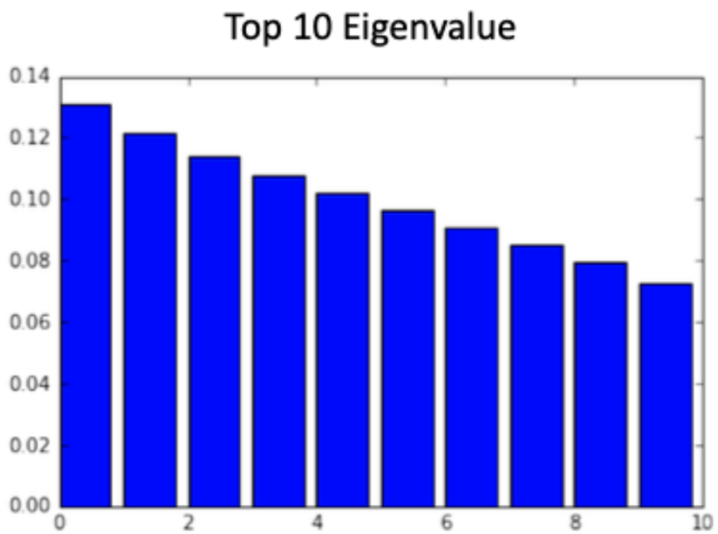
he: she $\approx$ professor: ? associate professor

Identify Gender Bias Directions (Space)



PCA ("he"- "she", "father"- "mother",...)

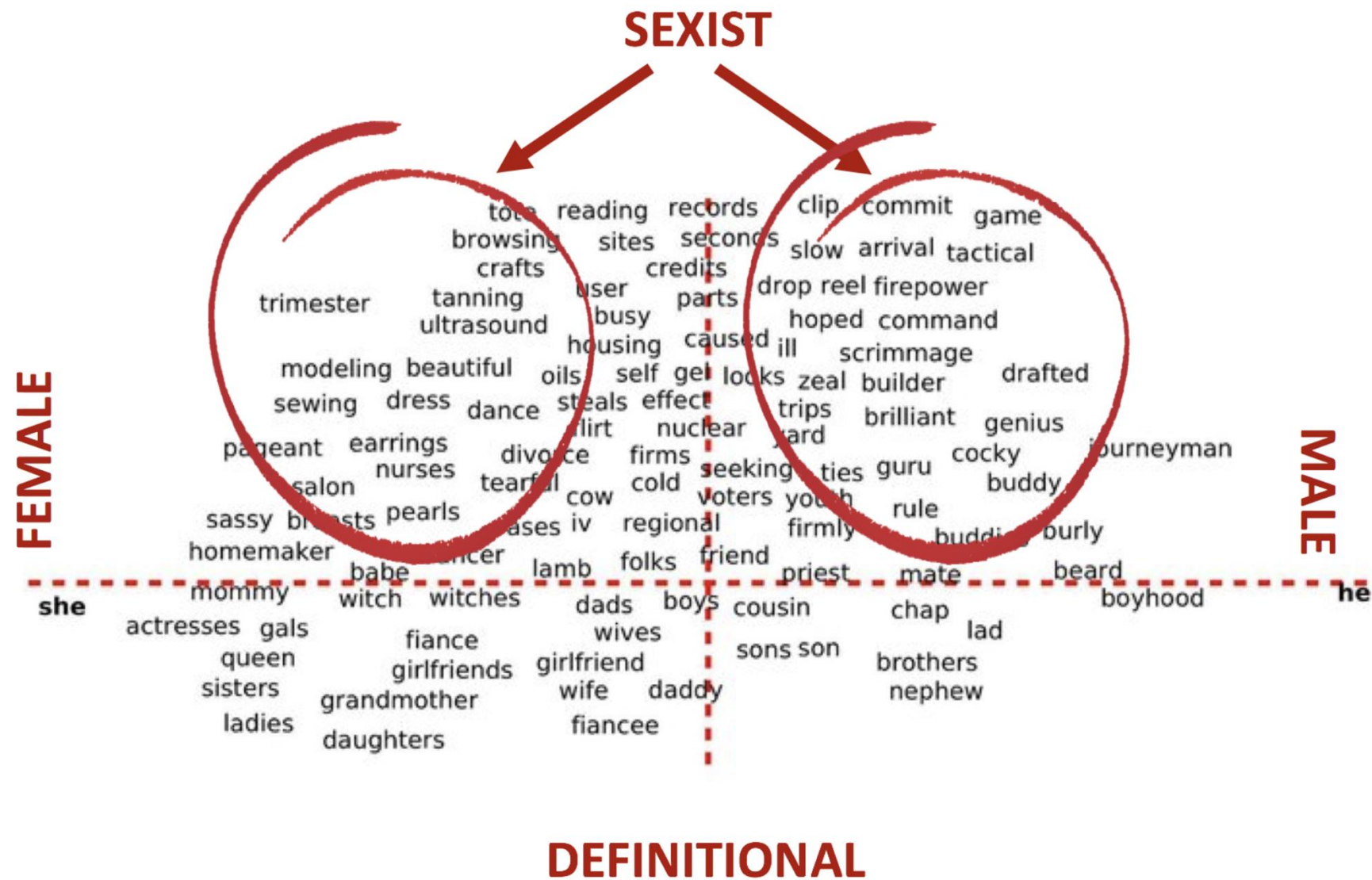
Gender Pair



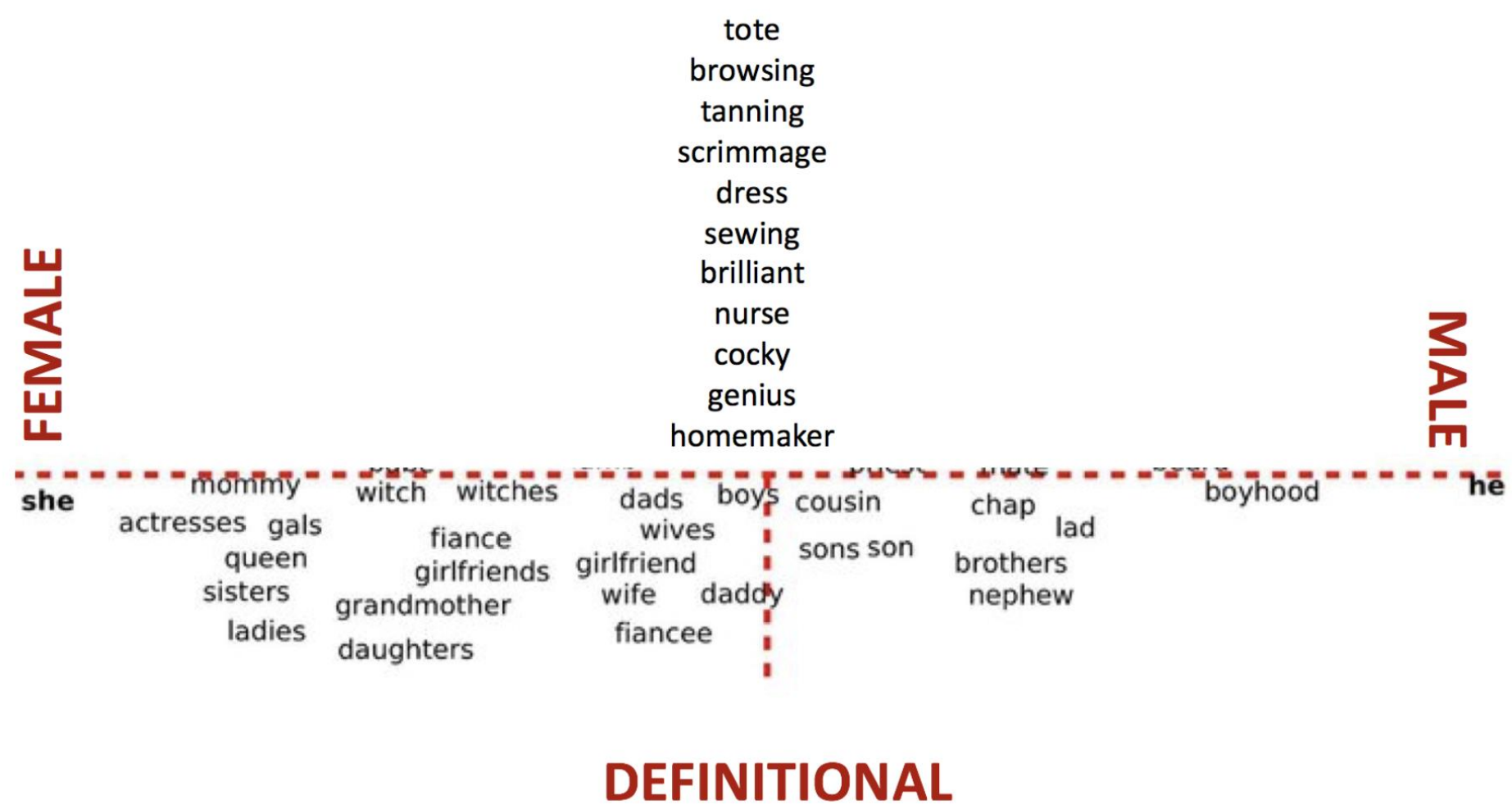
PCA ("dog"- "cat", "house"- "building",...)

Random Pair

Identify Gender Bias Directions (Space)



Debias with Projection



Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods

Jieyu Zhao[§] Tianlu Wang[†] Mark Yatskar[‡]
Vicente Ordonez[†] Kai-Wei Chang[§]

[§]University of California, Los Angeles {jyzhao, kwchang}@cs.ucla.edu

[†]University of Virginia {tw8bc, vicente}@virginia.edu

[‡]Allen Institute for Artificial Intelligence marky@allenai.org

Winograd Schema Challenge

- A test of a system's ability to perform commonsense reasoning
 - The trophy doesn't fit in the suitcase because it is too big
 - Anna didn't pass the message to Jessica because she was in a hurry
 - Frank felt threatened by Douglas because he was very competitive
 - The city council denied the protesters a permit because they feared violence.

WinoBias

Type 1

The physician hired the secretary because he was overwhelmed with clients.

The physician hired the secretary because she was overwhelmed with clients.

The physician hired the secretary because she was highly recommended.

The physician hired the secretary because he was highly recommended.

Type 2

The secretary called the physician and told him about a new patient.

The secretary called the physician and told her about a new patient.

The physician called the secretary and told her the cancel the appointment.

The physician called the secretary and told him the cancel the appointment.

Broaden the Vision: Geo-Diverse Visual Commonsense Reasoning

Da Yin Liunian Harold Li Ziniu Hu Nanyun Peng Kai-Wei Chang

Computer Science Department, University of California, Los Angeles

{da.yin, liunian.harold.li, bull, violetpeng, kwchang}@cs.ucla.edu

GIVL: Improving Geographical Inclusivity of Vision-Language Models with Pre-Training Methods

Da Yin¹ Feng Gao² Govind Thattai² Michael Johnston² Kai-Wei Chang^{1,2}

¹ University of California, Los Angeles ² Amazon Alexa AI

{da.yin, kwchang}@cs.ucla.edu, {fenggo, thattg, mjohnstn}@amazon.com

Evaluation for Cultural Bias

Diverse Scenarios in Different Regions



Africa



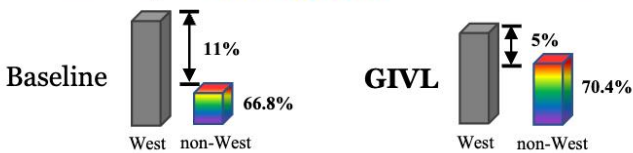
Western



East Asia



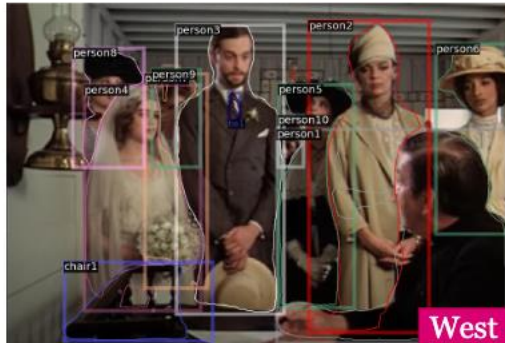
Performance Gap of non-Western and Western Scenarios



South Asia

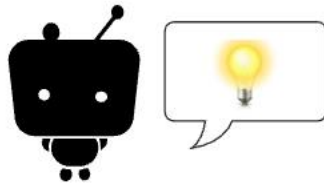


Evaluation for Cultural Bias



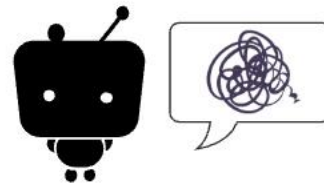
Question: What are [person3] and [person4] participating in?

- A.
- B. They are in a wedding.
- C.
- D.



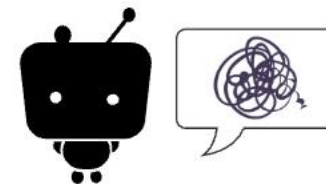
Question: What are [person1] and [person2] participating in?

- A.
- B. They are in a wedding.
- C.
- D.



Question: What are [person1] and [person2] participating in?

- A.
- B. They are in a wedding.
- C.
- D.



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

**Jieyu Zhao[§] Tianlu Wang[§] Mark Yatskar[‡]
Vicente Ordonez[§] Kai-Wei Chang[§]**

[§]University of Virginia
{jz4fu, tw8cb, vicente, kc2wc}@virginia.edu

[‡]University of Washington
my89@cs.washington.edu

Visual Semantic Role Labeling



COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	PASTA
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN



COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	FRUIT
HEAT	∅
TOOL	KNIFE
PLACE	KITCHEN



COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	MEAT
HEAT	STOVE
TOOL	SPATULA
PLACE	OUTSIDE



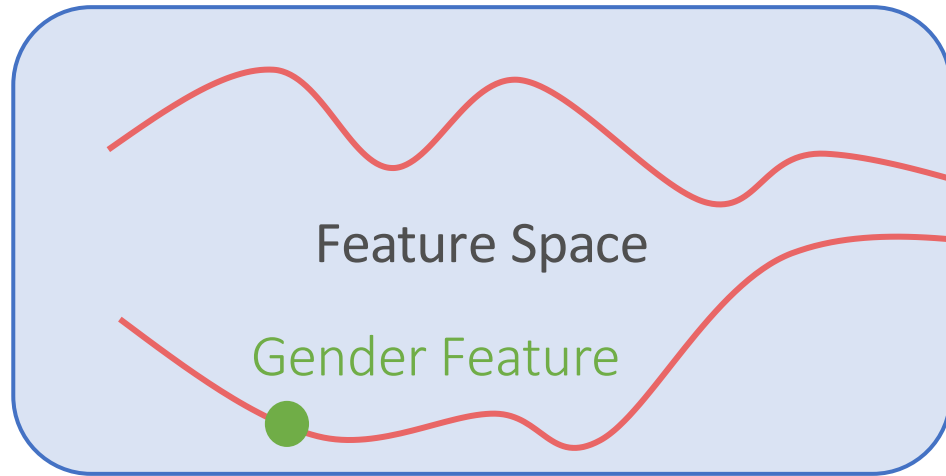
COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN



COOKING	
ROLE	VALUE
AGENT	MAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

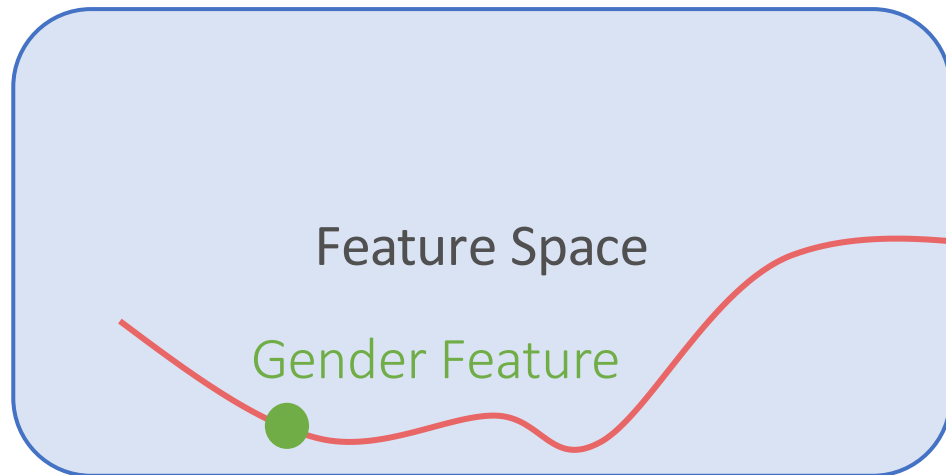
Recap: Bias or Features?

My Explanation



Prediction

If other neutral features exist,
don't use sensitive features



Prediction

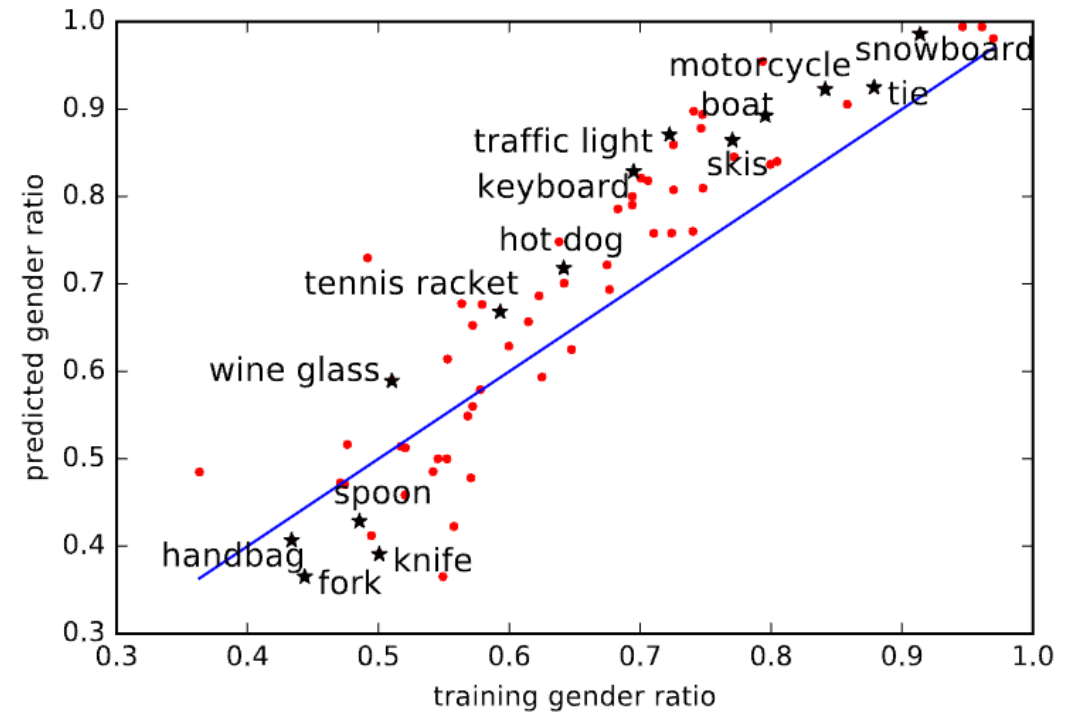
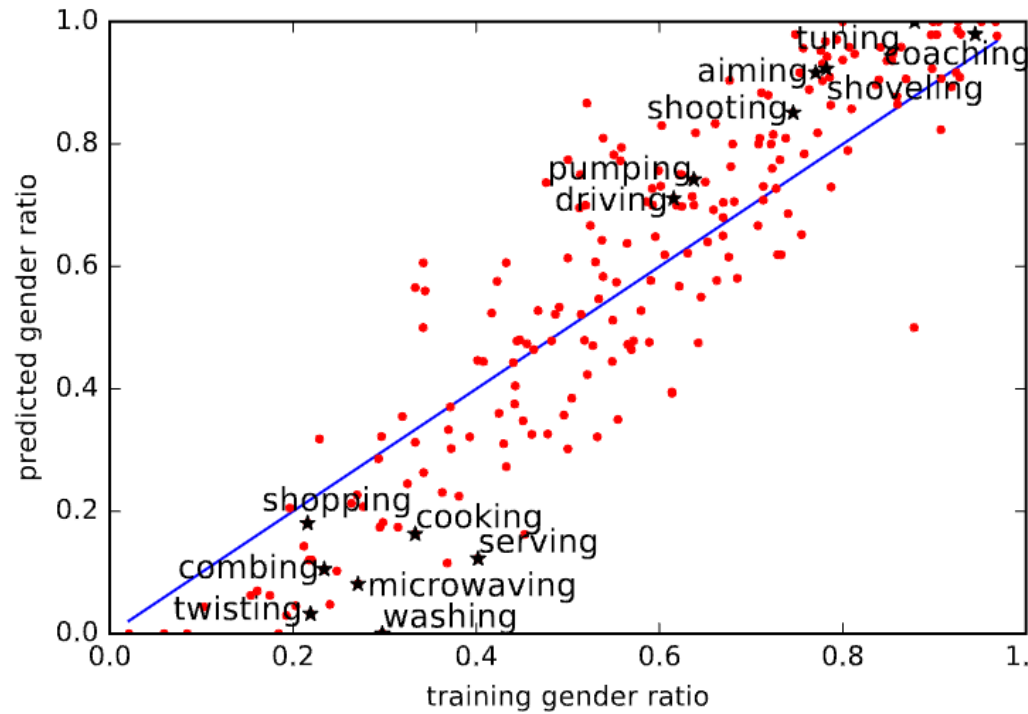
If no other neutral features,
no amplification is allowed

70% male and 30% female

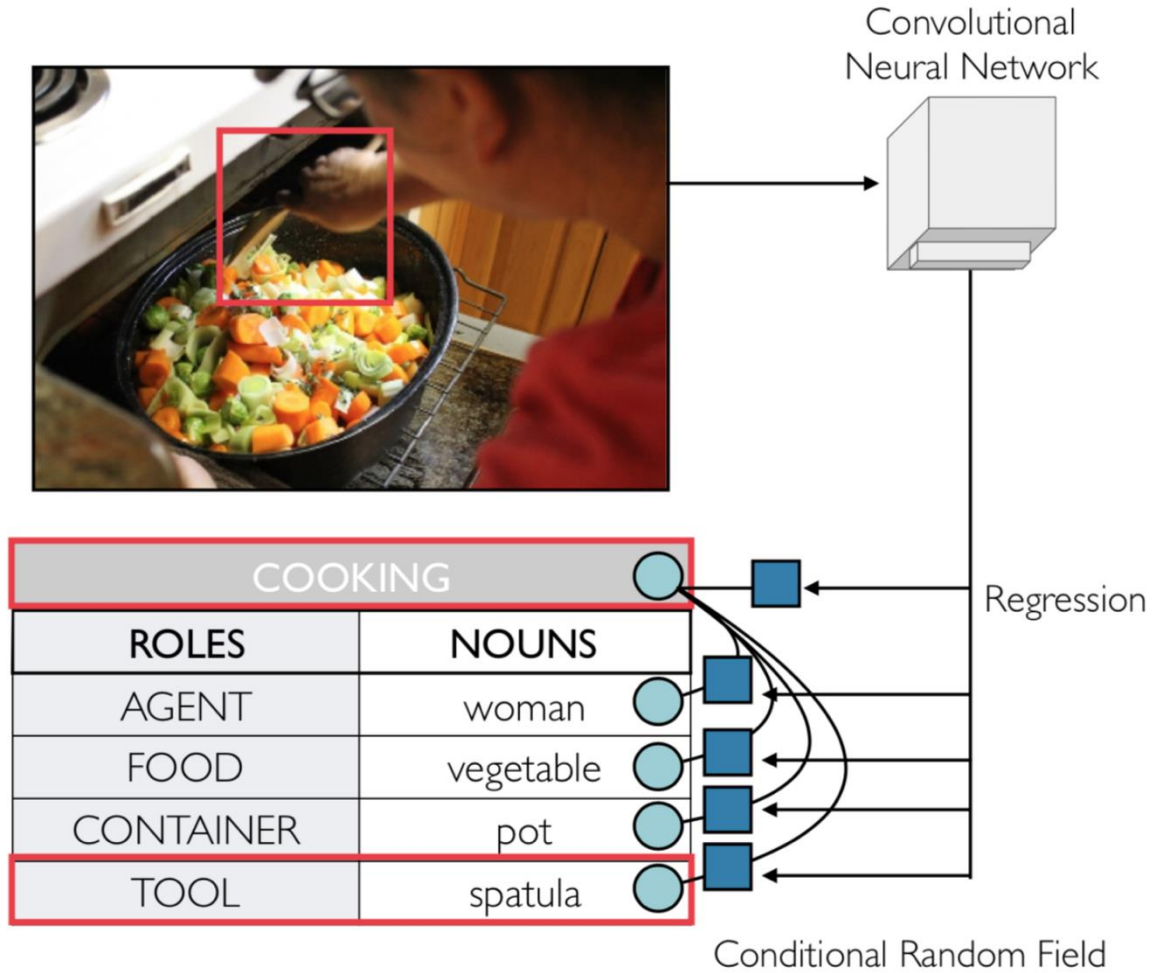
$P(Y \mid \text{male}) = 70\%$

Bias Amplification

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



Structured Prediction Problem



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

Corpus-Level Constraints

Integer Linear Program

$$\sum_i \max_{y_i} s(y_i, \text{image})$$

Goal of the original model

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

Our control for calibration

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

Lagrangian Relaxation

Integer Linear Program

$$\sum_i \max_{y_i} s(y_i, \text{image})$$

Goal of the original model

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

Our control for calibration

$$\max_{\{y^i\} \in \{Y^i\}} \sum_i f_{\theta}(y^i, i), \quad \text{s.t.} \quad A \sum_i y^i - b \leq 0$$

Lagrangian : $\sum_i f_{\theta}(y^i) - \sum_{j=1}^l \lambda_j (A_j \sum_i y^i - b_j) \quad \lambda_j \geq 0$

Lagrangian Relaxation

$$\max_{\{y^i\} \in \{Y^i\}} \sum_i f_\theta(y^i, i), \quad \text{s.t.} \quad A \sum_i y^i - b \leq 0$$

Lagrangian : $\sum_i f_\theta(y^i) - \sum_{j=1}^l \lambda_j (A_j \sum_i y^i - b_j) \quad \lambda_j \geq 0$

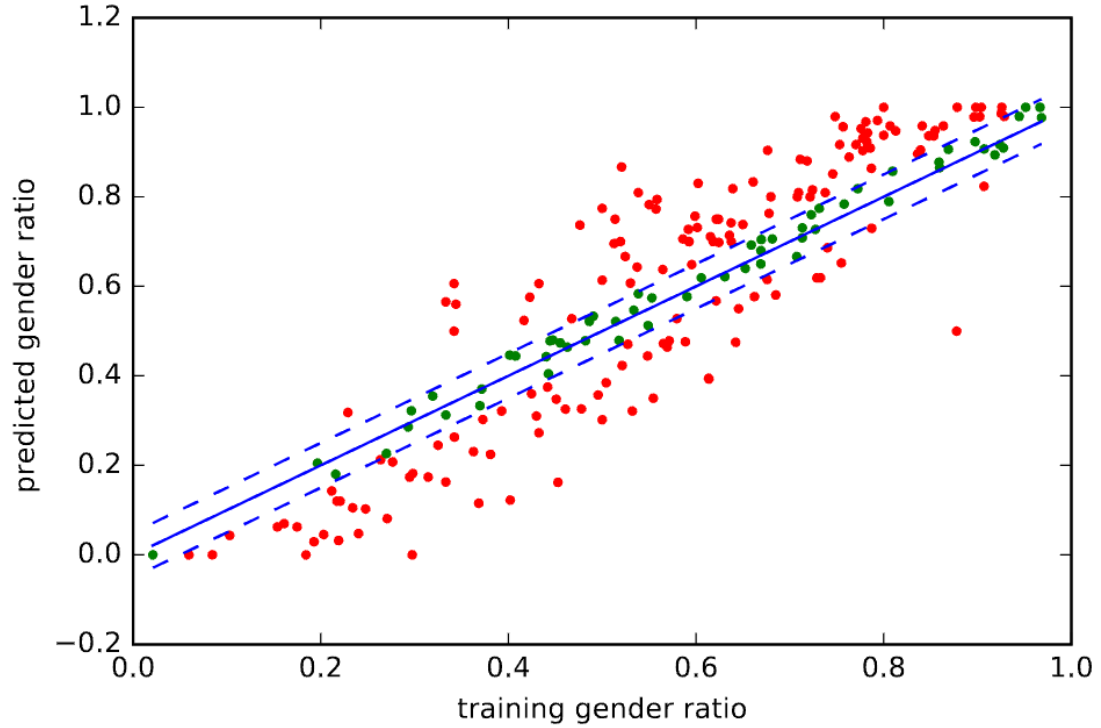
- 1) At iteration t , get the output solution of each instance i

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

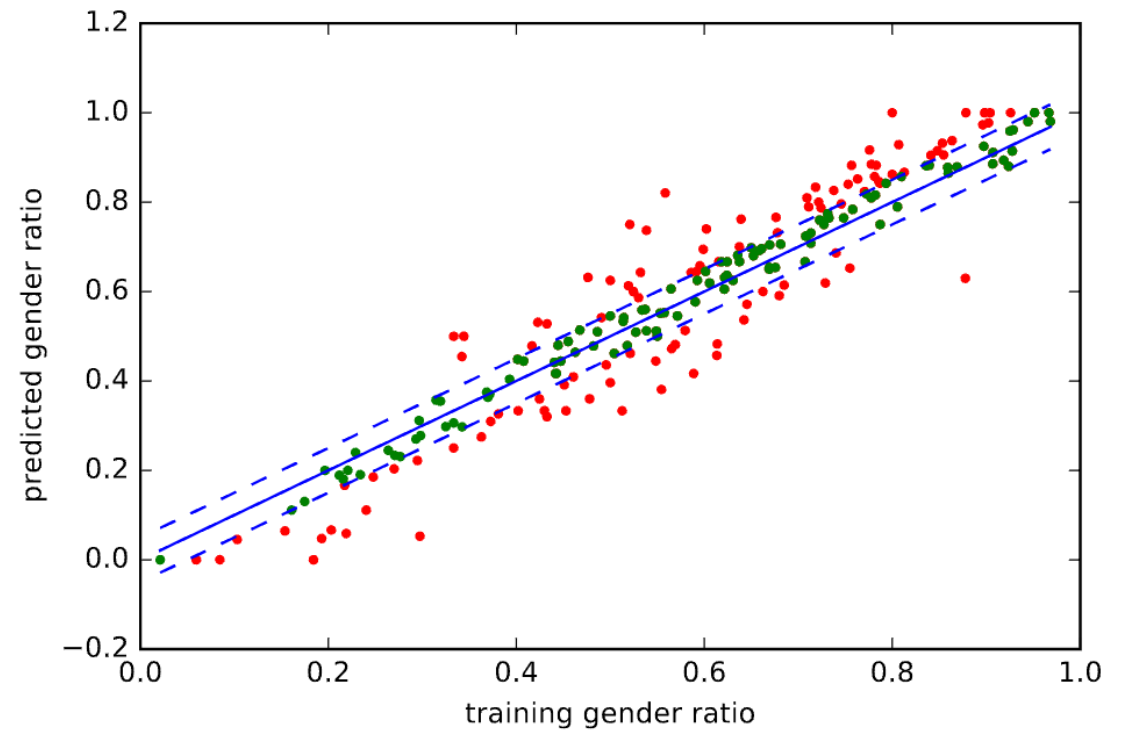
- 2) update the Lagrangian multipliers.

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

Results



(a) Bias analysis on imSitu vSRL without RBA



(c) Bias analysis on imSitu vSRL with RBA

Results

Mitigating Gender Bias Amplification in Distribution by Posterior Regularization

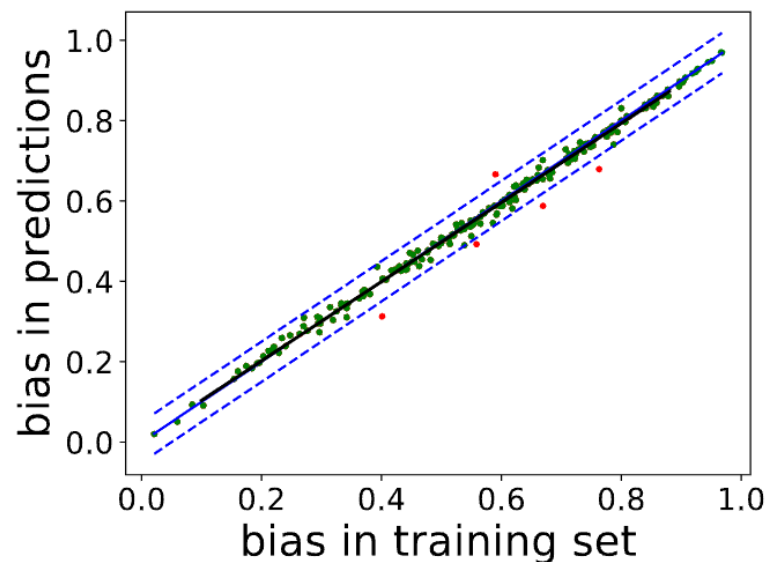
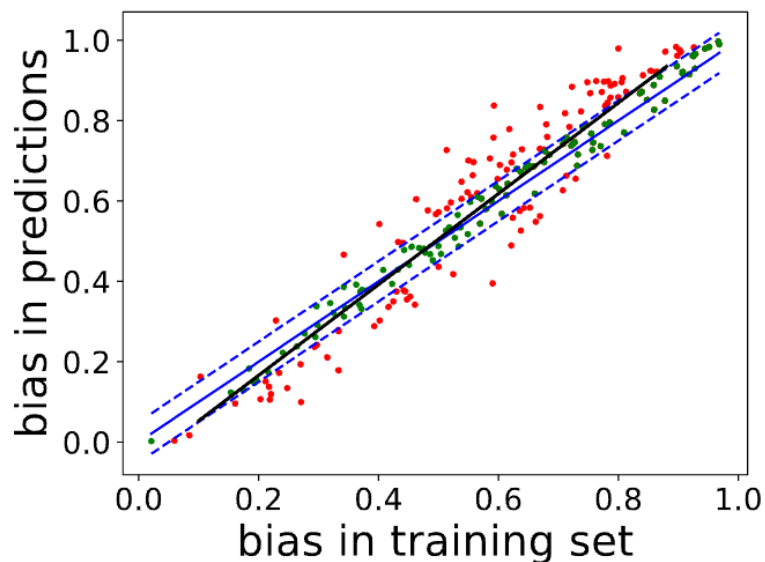
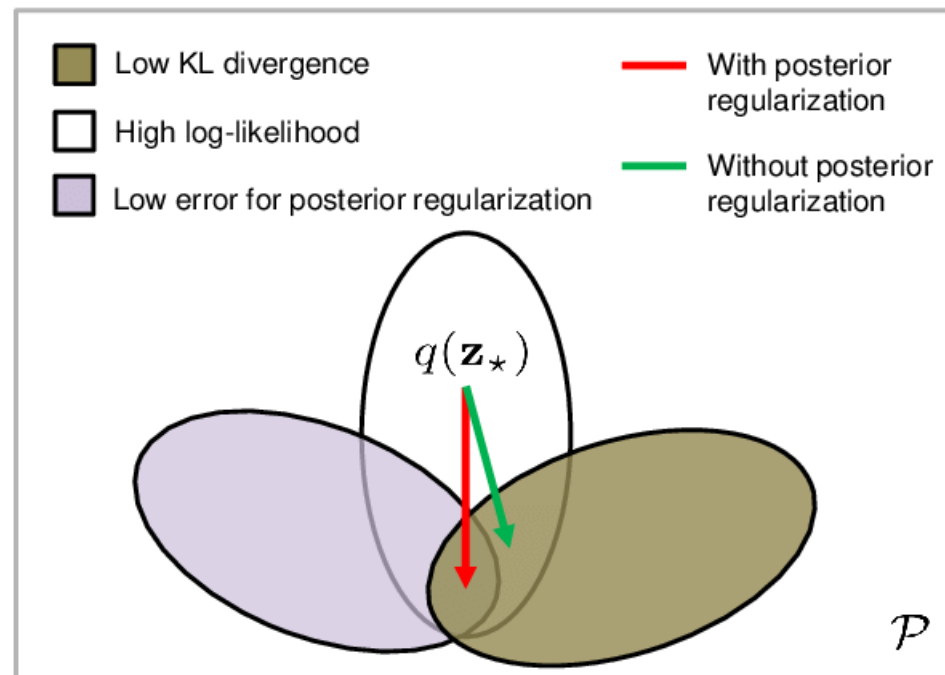
Shengyu Jia^{♣,*}, Tao Meng^{♣,*}, Jieyu Zhao[♣], Kai-Wei Chang[♣]

[♣] Tsinghua University

[♣] University of California, Los Angeles

jiasyl6@mails.tsinghua.edu.cn,

{mengt18, jieyuzhao, kwchang}@ucla.edu



DISTRIBUTIONALLY ROBUST NEURAL NETWORKS FOR GROUP SHIFTS: ON THE IMPORTANCE OF REGULARIZATION FOR WORST-CASE GENERALIZATION

Shiori Sagawa*

Stanford University

ssagawa@cs.stanford.edu

Pang Wei Koh*

Stanford University

pangwei@cs.stanford.edu

Tatsunori B. Hashimoto

Microsoft

tahashim@microsoft.com

Percy Liang

Stanford University

pliang@cs.stanford.edu

Spurious Correlations

Common training examples

Test examples

Waterbirds

y: waterbird
a: water
background



y: landbird
a: land
background



y: waterbird
a: land
background

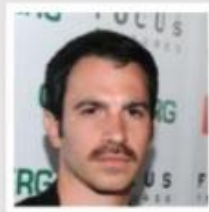


CelebA

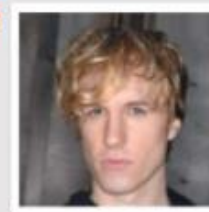
y: blond hair
a: female



y: dark hair
a: male



y: blond hair
a: male



MultiNLI

y: contradiction
a: has negation
(P) The economy could be still better.
(H) The economy has never been better.

y: entailment
a: no negation
(P) Read for Slate's take on Jackson's findings.
(H) Slate had an opinion on Jackson's findings.

y: entailment
a: has negation
(P) There was silence for a moment.
(H) There was a short period of time where no one spoke.

Group Distributionally Robust Optimization (Group DPO)

Standard Optimization

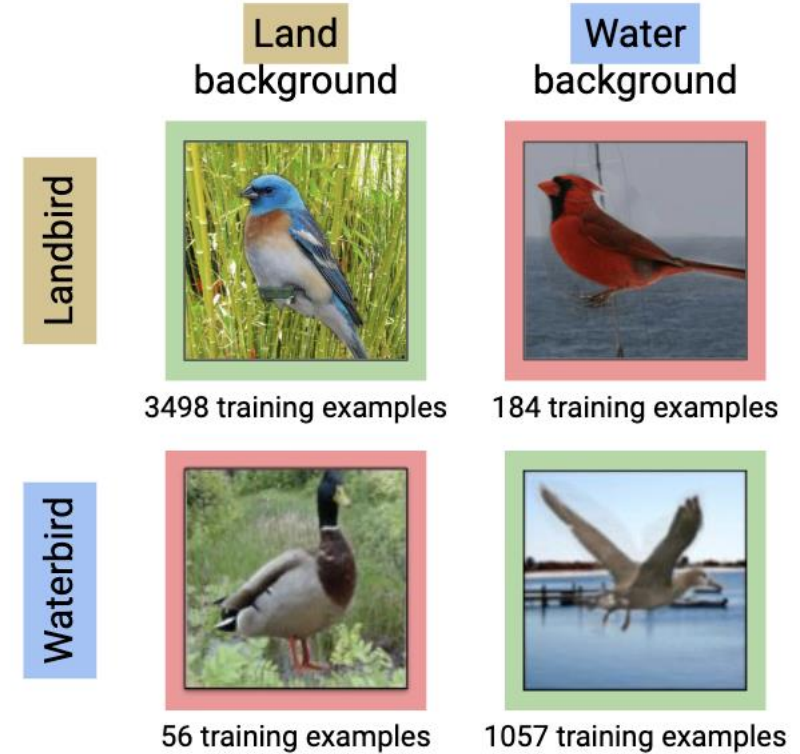
$$\hat{\theta}_{\text{ERM}} := \arg \min_{\theta \in \Theta} \mathbb{E}_{(x,y) \sim \hat{P}} [\ell(\theta; (x, y))].$$

Worst Case Optimization

$$\min_{\theta \in \Theta} \left\{ \mathcal{R}(\theta) := \sup_{Q \in \mathcal{Q}} \mathbb{E}_{(x,y) \sim Q} [\ell(\theta; (x, y))] \right\}$$

Worst Group Optimization

$$\hat{\theta}_{\text{DRO}} := \arg \min_{\theta \in \Theta} \left\{ \hat{\mathcal{R}}(\theta) := \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y) \sim \hat{P}_g} [\ell(\theta; (x, y))] \right\}$$



Just Train Twice: Improving Group Robustness without Training Group Information

Evan Zheran Liu^{*1} Behzad Haghgoo^{*1} Annie S. Chen^{*1} Aditi Raghunathan¹ Pang Wei Koh¹
Shiori Sagawa¹ Percy Liang¹ Chelsea Finn¹

BLIND: Bias Removal With No Demographics

Hadas Orgad **Yonatan Belinkov^{*}**
orgad.hadas@cs.technion.ac.il belinkov@technion.ac.il
Technion – Israel Institute of Technology

Just Train Twice

Previous work needs to know spurious features in advance

Identify not easy-to-learn spurious correlations

$$E = \{(x_i, y_i) \text{ s.t. } \hat{f}_{\text{id}}(x_i) \neq y_i\}$$

Upweight hard examples

$$J_{\text{up-ERM}}(\theta, E) = \left(\lambda_{\text{up}} \sum_{(x,y) \in E} \ell(x, y; \theta) + \sum_{(x,y) \notin E} \ell(x, y; \theta) \right)$$

Method	Group labels in train set?	Waterbirds		CelebA		MultiNLI		CivilComments-WILDS	
		Avg Acc.	Worst-group Acc.	Avg Acc.	Worst-group Acc.	Avg Acc.	Worst-group Acc.	Avg Acc.	Worst-group Acc.
ERM	No	97.3%	72.6%	95.6%	47.2%	82.4%	67.9%	92.6%	57.4%
JTT (Ours)	No	93.3%	86.7%	88.0%	81.1%	78.6%	72.6%	91.1%	69.3%
Group DRO (Sagawa et al., 2020a)	Yes	93.5%	91.4%	92.9%	88.9%	81.4%	77.7%	88.9%	69.9%

Debias With No Demographics

BLIND: Bias Removal With No Demographics

Hadas Orgad

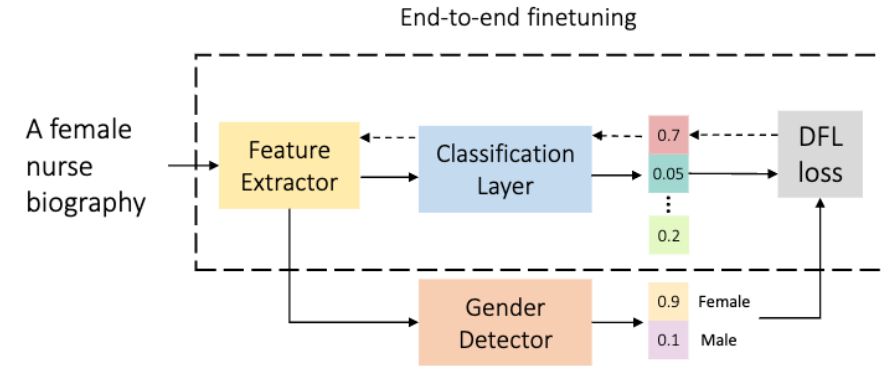
orgad.hadas@cs.technion.ac.il

Yonatan Belinkov*

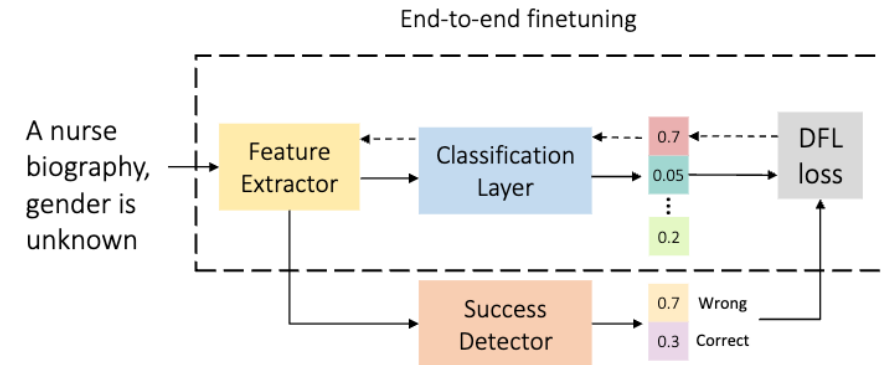
belinkov@technion.ac.il

Technion – Israel Institute of Technology

Previous work needs to know spurious features in advance



(a) With demographic annotations. Demographics detector learns to predict the demographic data, e.g., gender.



(b) BLIND: Without demographic annotations. Success detector learns to predict when the main model is correct. Supervision is based only on the downstream task labels.

Questions?