Fair Visual Representations

May 8, 2020

Dr. Wei Wei, Prof. James Landay

CS 335: Fair, Accountable, and Transparent (FAccT) Deep Learning

Stanford University

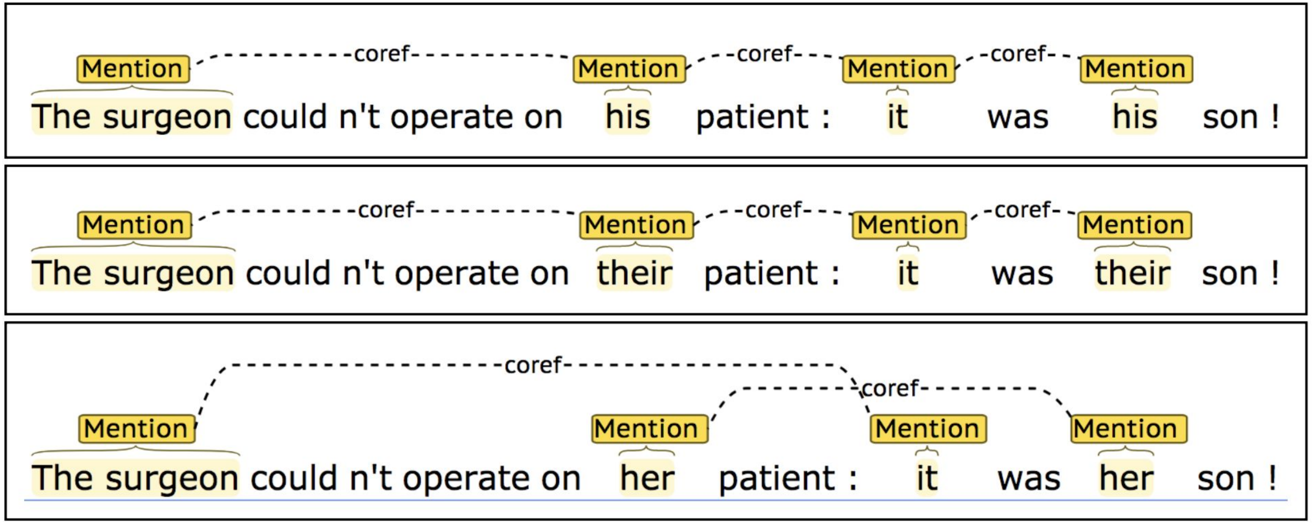
Recap

- Gender Swapping for Coreference Resolution

Original sample

Gender swap

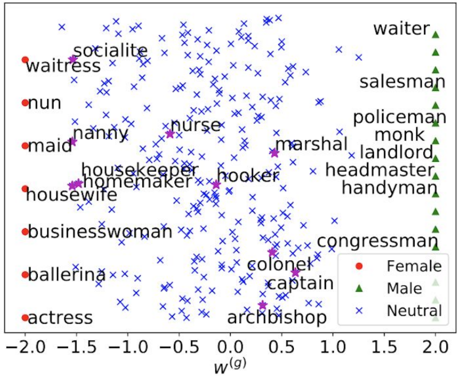
Gender swap



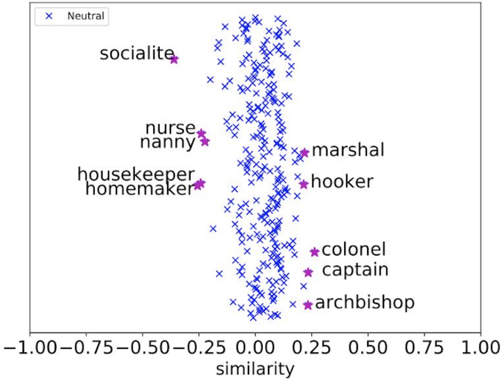
[Rudinger et al, 2018](#)

Recap

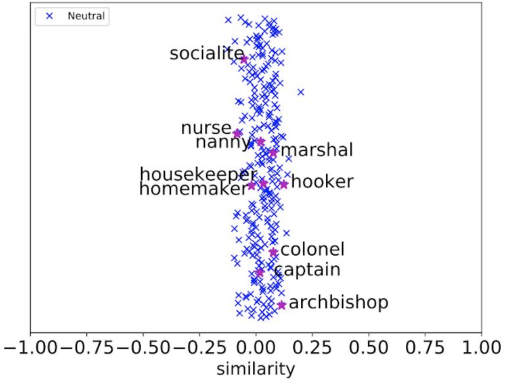
- Debiasing Word Embedding by Gender Attribute Separation



$w^{(g)}$ of All Occupations



$w^{(a)}$ of GloVe for Gender Neutral Occupations

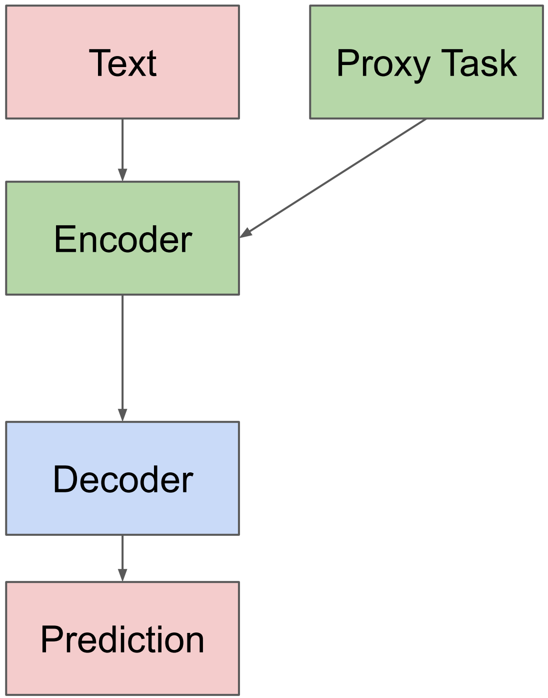


$w^{(a)}$ of Gender-Neutral GloVE for Gender Neutral Occupations

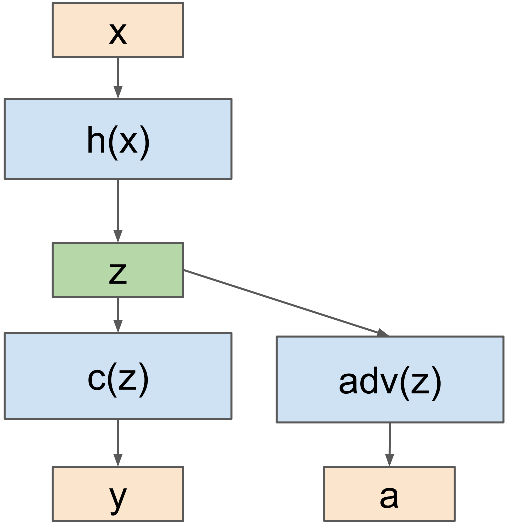
$w^{(g)}$ - Gender-related Components
$w^{(a)}$ - Gender-neutral Components

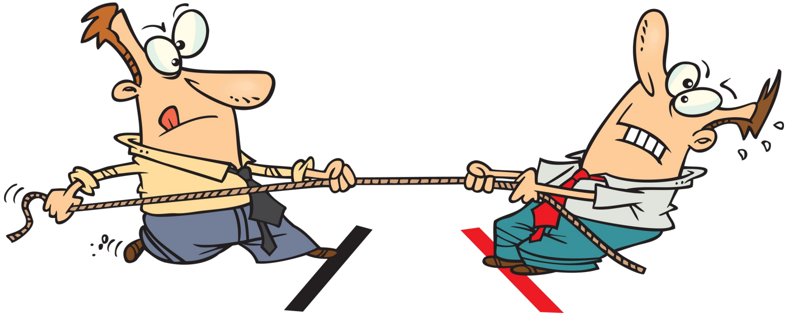
The Use of Pre-trained NLP Encoders

- Pre-trained Encoders Are Widely Used in NLP
  - Transfer information from a related domain
  - Boost performance on a small data set
  - Trained through a proxy task
- Pre-trained NLP Encoders
  - ELMO (Peters et al 2018)
  - BERT (Devlin et al, 2018)
  - XLNet (Yang et al, 2019)
- Can Pre-trained Encoders Be Biased?



Adversarial Learning





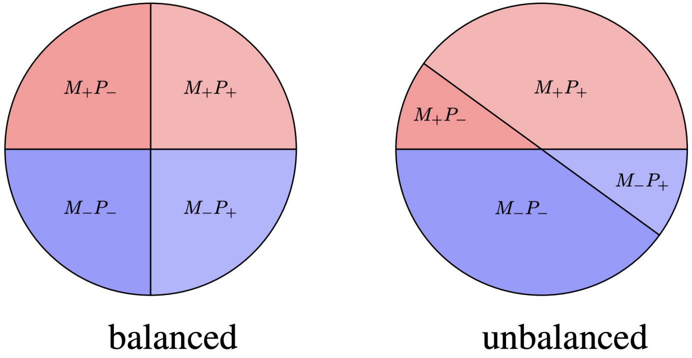
Predicting y

Reconstruct a

[Elazar et al, 2018](#)

Twitter Prediction Problem

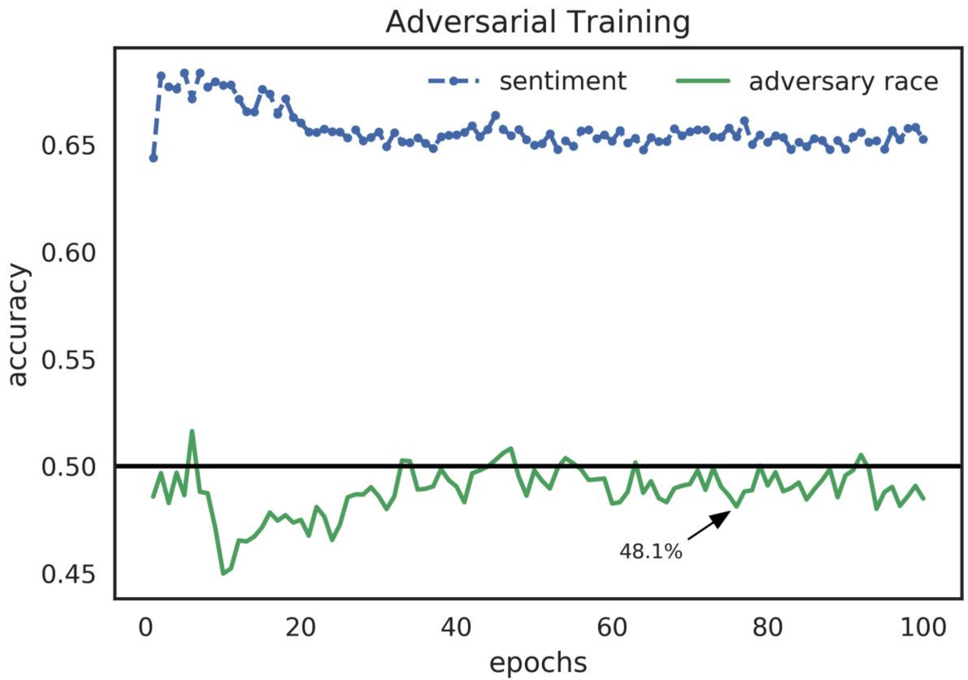
- Twitter Sentiment & Mention Detection
- Protected Attributes
  - Race
  - Gender
  - Age
- Leakage
  - Predict protected attributes



Data	Task	Protected Attribute	Balanced Task Acc	Balanced Leakage	Unbalanced Task Acc	Unbalanced Leakage
DIAL	Sentiment	Race	67.4	64.5	79.5	73.5
	Mention	Race	81.2	71.5	86.0	73.8
PAN16	Mention	Gender	77.5	60.1	76.8	64.0
		Age	74.7	59.4	77.5	59.7

Elazar et al, 2018

Main Task and Adversary Accuracies



[Elazar et al, 2018]()

Beefing Up the Adversary

- Increase the Capacity of the Adversary
  - Model Capacity
  - Weight on Loss
  - Ensemble

		DIAL			PAN16					
Method	Parameter	Sentiment	Race	Δ	Mention	Gender	Δ	Mention	Age	Δ
No Adversary Baseline	-	67.4	14.5	-	77.5	10.1	-	74.7	9.4	-
Standard Adversary	(300/1.0/1)	64.7	6.0	5.0	75.6	8.5	8.0	72.5	7.3	6.9
Adv-Capacity	500	64.1	6.7	5.2	73.8	8.1	6.7	71.4	**4.3**	4.1
	1000	63.4	7.1	4.9	75.2	8.9	7.0	71.6	6.3	4.0
	2000	65.2	8.1	6.9	76.1	6.7	6.4	71.9	6.0	5.7
	5000	63.9	**6.2**	3.7	74.5	5.6	1.6	73.0	10.2	9.6
	8000	65.0	7.1	4.8	75.7	**5.4**	4.2	71.9	9.8	7.3
λ	0.5	63.9	**6.8**	6.2	75.6	7.8	6.8	73.1	4.8	3.4
	1.5	64.9	7.4	5.4	75.6	**4.9**	2.4	72.5	6.8	5.8
	2.0	64.2	7.3	5.9	76.0	-7.2	6.7	72.1	8.5	7.7
	3.0	65.8	10.2	10.1	73.7	6.4	6.1	72.5	-6.3	5.2
	5.0	50.0	-	-	73.6	6.5	5.7	69.0	**3.2**	2.9
Ensemble	2	62.4	7.4	5.4	74.8	6.4	5.0	72.8	8.8	8.3
	3	66.5	6.5	5.0	75.3	4.9	3.1	72.1	6.7	6.0
	5	63.8	**4.8**	2.6	74.3	**4.1**	3.0	70.1	**5.7**	5.4

Δ - the difference between the attacker score and the corresponding adversary's accuracy

Elazar et al, 2018

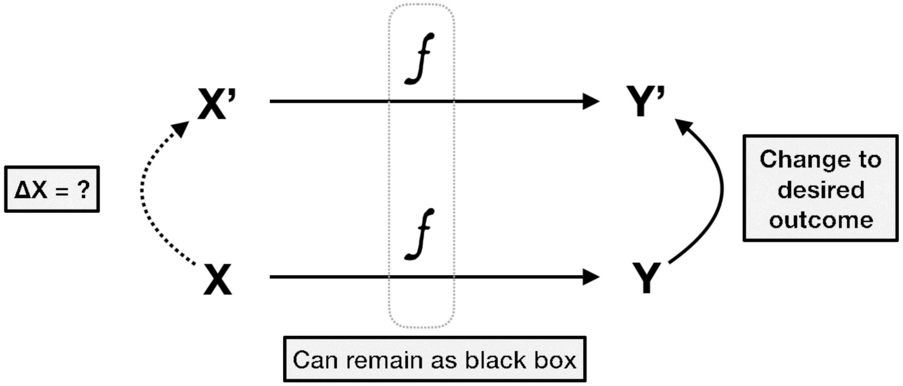
Outline

- Counterfactual Fairness
- Counterfactual Face Attribution
- Gender Equalized Image Captioning
- Adversarial Removal of Gender Features

Counterfactual Explanation

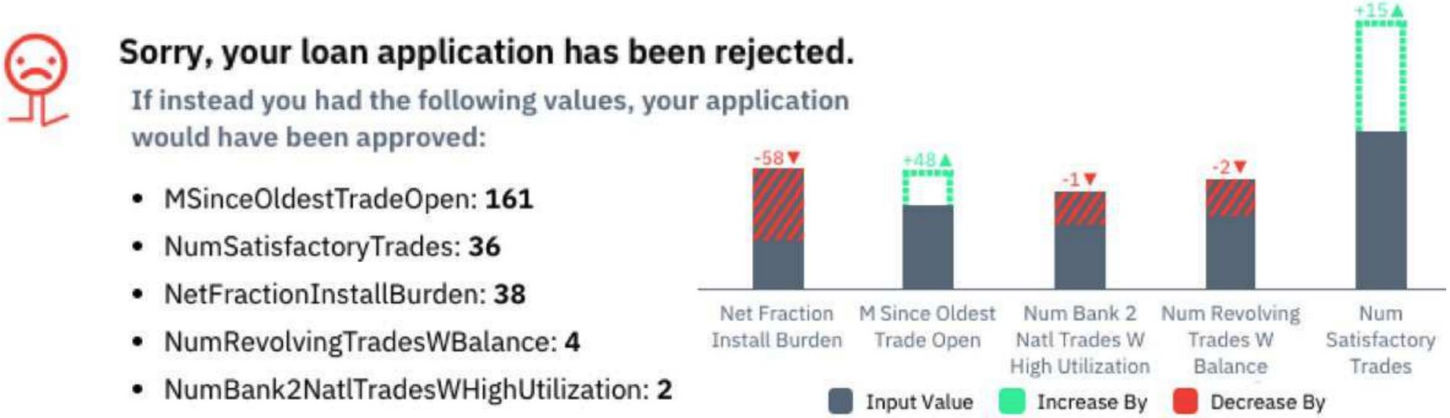
$$\underbrace{x'}_{\text{counterfactual example}} = arg\min_{x'} \lambda(\hat{f}(x') - \underbrace{y'}_{\text{desired outcome}})^2 + \underbrace{d(x, x')}_{\text{distance function}}$$

Increase λ while $|\hat{f}(x') - y'| > \varepsilon$



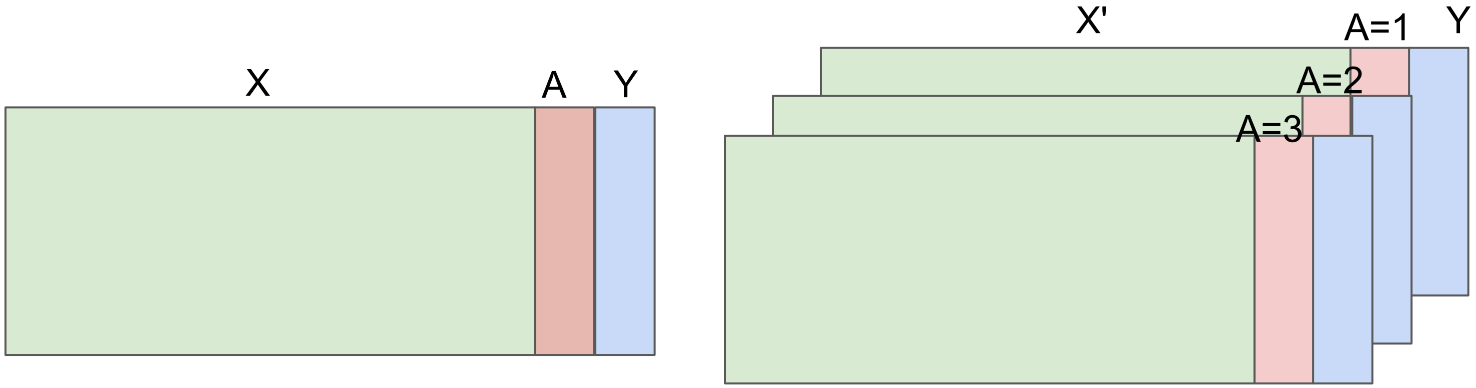
[Grath et al, 2018](#)

Counterfactual Explanations



Grath et al, 2018

Counterfactual Fairness



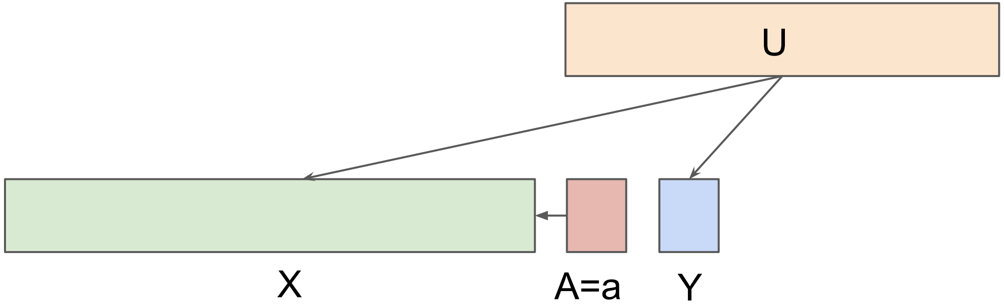
Real Examples

Counterfactual Examples

$$\underbrace{P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a)}_{\text{Real Examples}} = \underbrace{P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a)}_{\text{Counterfactual Examples}}$$

[Kusner et al, 2017]

Causal View of Counterfactual Fairness

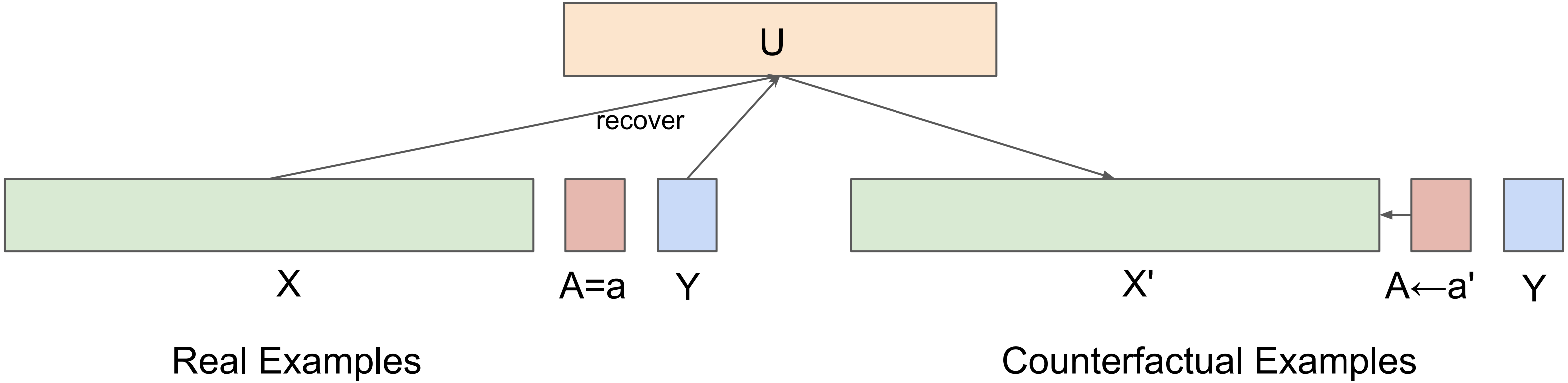


Real Examples

$$\underbrace{P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a)}_{\text{Real Examples}} = \underbrace{P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a)}_{\text{Counterfactual Examples}}$$

Kusner et al, 2017

Causal View of Counterfactual Fairness

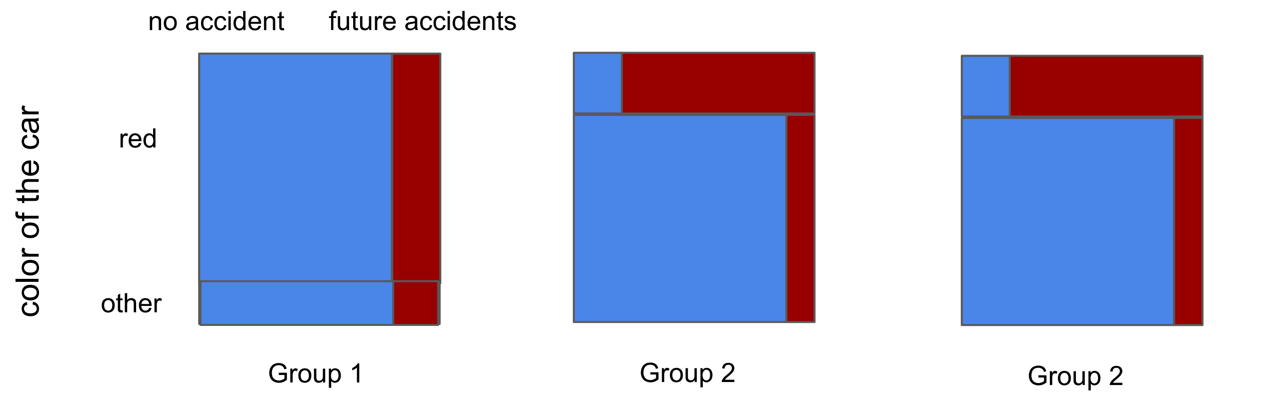


$$\underbrace{P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a)}_{\text{Real Examples}} = \underbrace{P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a)}_{\text{Counterfactual Examples}}$$

Kusner et al, 2017

Red Car Problem

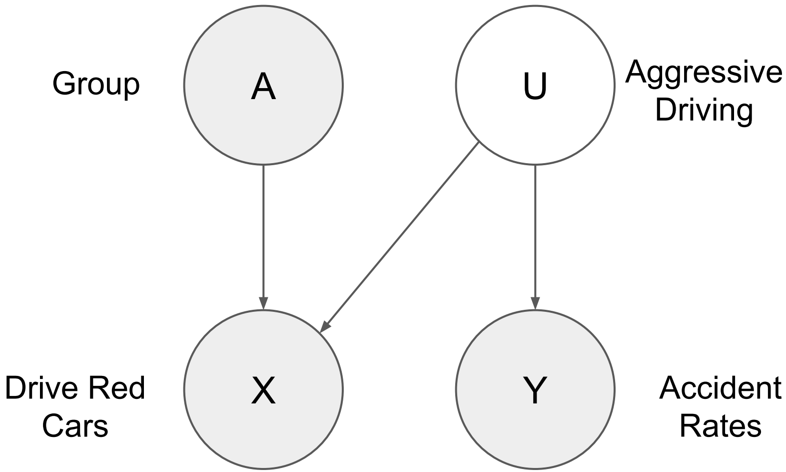
- Develop a Fair Algorithm to Determine Insurance Premium

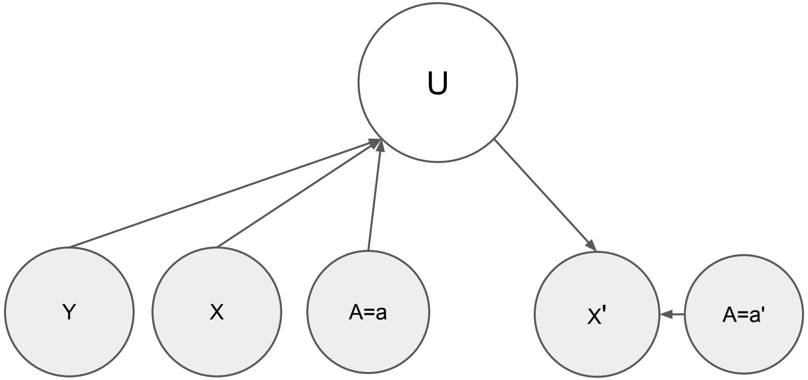


$P(y = acc \mid A = 1) = 0.2$ $P(y = acc \mid A = 2) = 0.2$ $P(y = acc \mid A = 3) = 0.2$

[Kusner et al, 2017]

Causal Perspective of the Red Car Problem





Observed Variables
Latent (Unobserved) Variables

$$P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a)$$

Kusner et al, 2017

Fairness Criteria

Individual Treatment	Group Treatment
Fairness Through Unawareness Excludes Sensitive Information A from the predictor	**Demographic Parity** $P(\hat{Y}=1 \mid A=1) = P(\hat{Y}=1 \mid A=0)$
Individual Fairness $M(x_i) \approx M(x_j) \mid d(x_i, x_j) \approx 0$	**Equal Opportunity/Odds** $P(\hat{Y}=1 \mid A=0, Y=1) = P(\hat{Y}=1 \mid A=1, Y=1)$ $P(\hat{Y}=1 \mid A=0, Y) = P(\hat{Y}=1 \mid A=1, Y)$
Counterfactual Fairness $P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a)$	

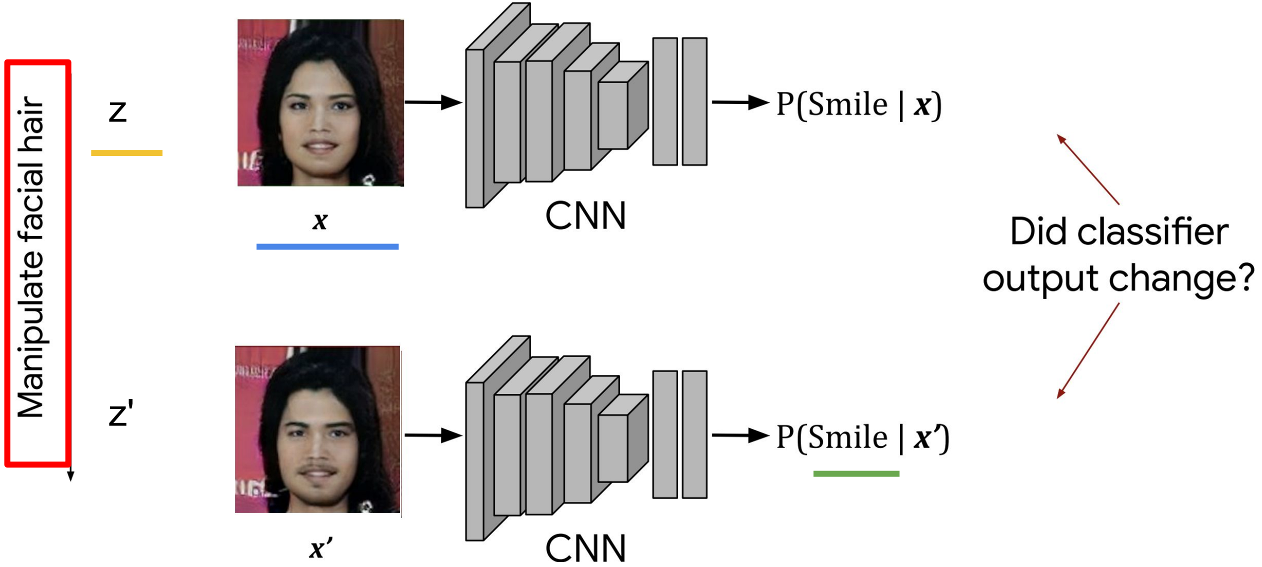
$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a)$$

Outline

- Counterfactual Fairness
- Counterfactual Face Attribution
- Gender Equalized Image Captioning
- Adversarial Removal of Gender Features

Counterfactual Face Attribution

- Evaluate the Counterfactual Fairness of Face Recognition Systems

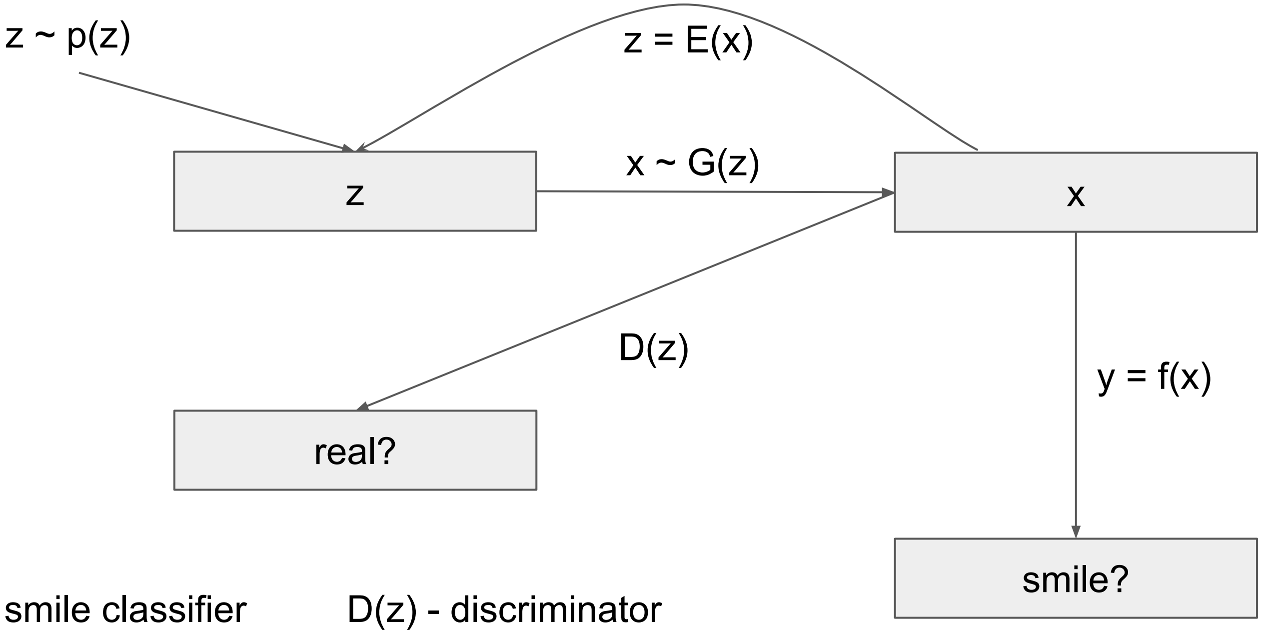


$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a)$$

CelebA Dataset

[Liu et al, 2015]

Model Architecture

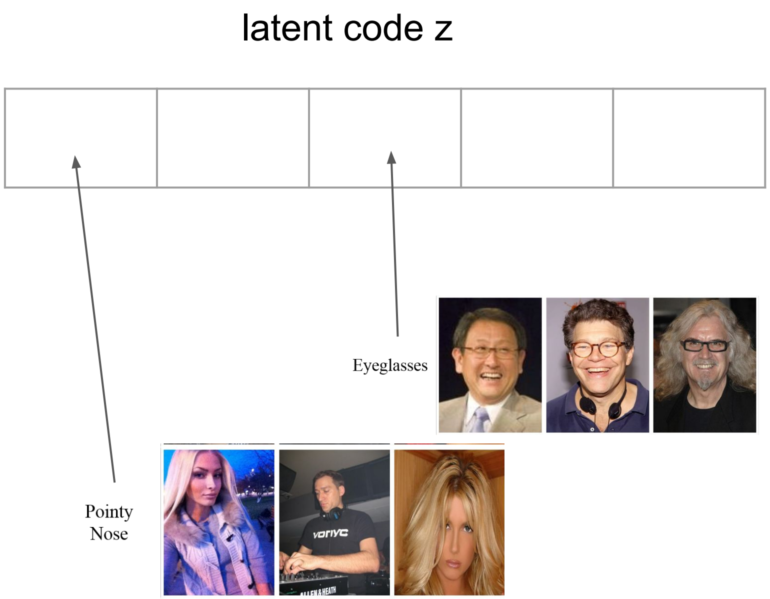


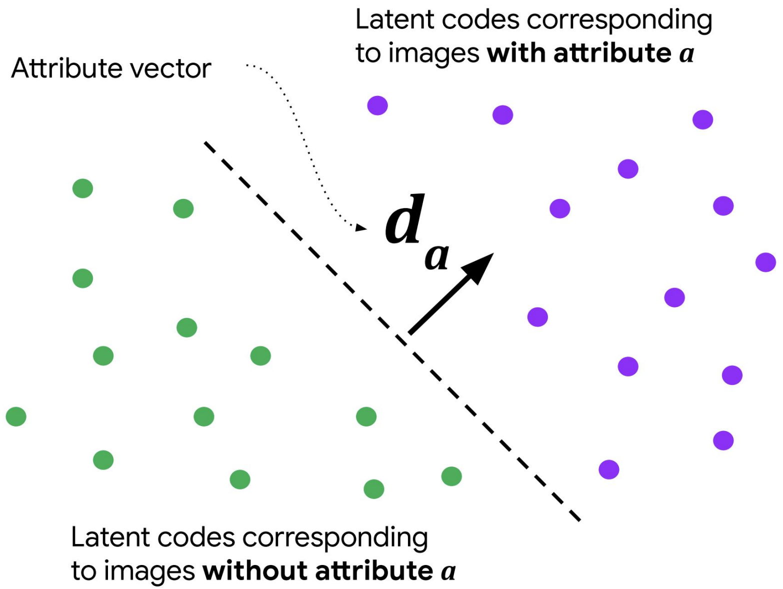
f(x) - smile classifier
G(z) - generator

D(z) - discriminator
E (x) - encoder

Denton et al, 2019

Latent Code Attribution

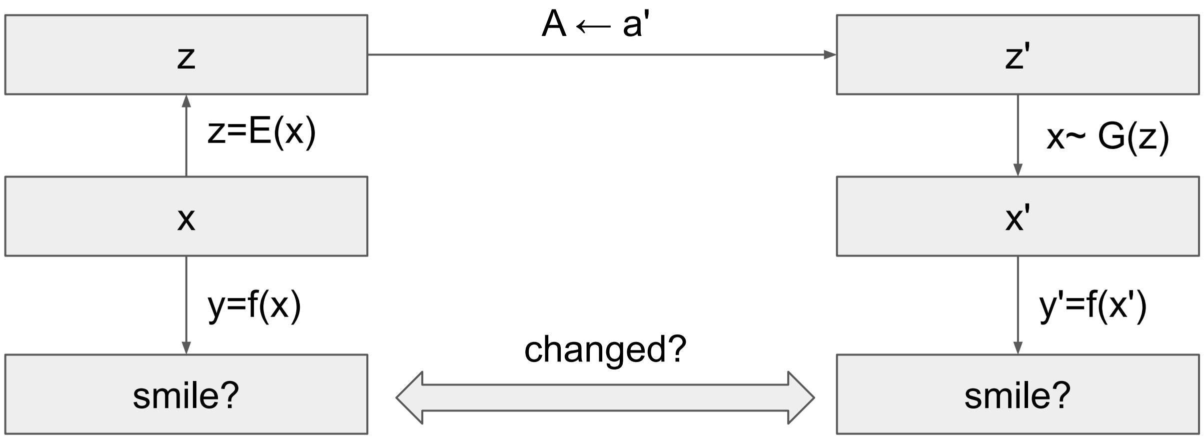




Latent Code Manipulation

[Denton et al, 2019]

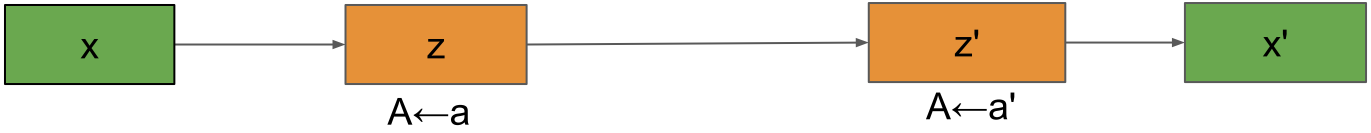
Counterfactual Fairness Assessment



$$P(\hat{Y}_{A\leftarrow a}\,(U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a)$$

Denton et al, 2019

Sensitivity



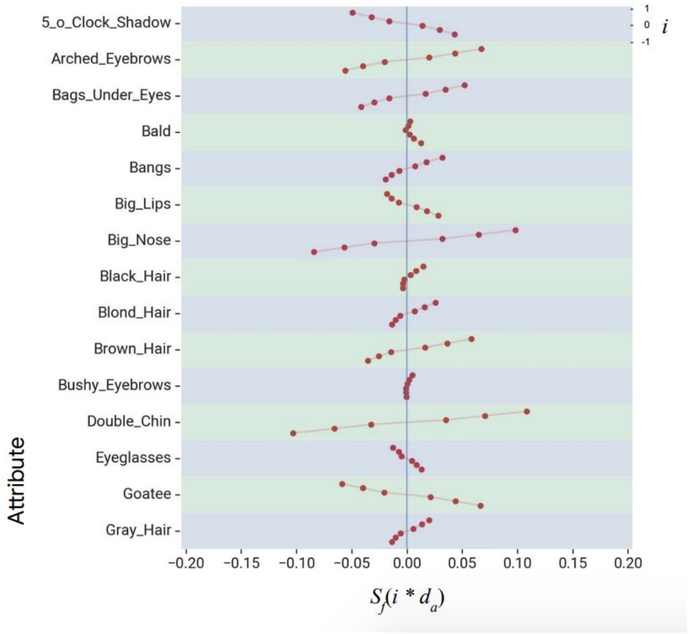
$$S_f(d) = \mathbb{E}_{z \sim p(z)}[f(G(z+d)) - f(G(z))]$$

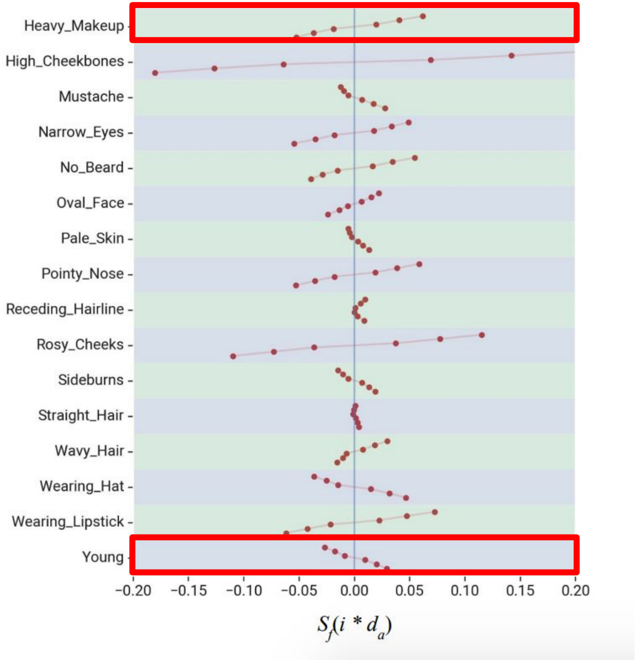
f(x) - smile classifier
G(z) - generator
D(z) - discriminator

$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a)$$

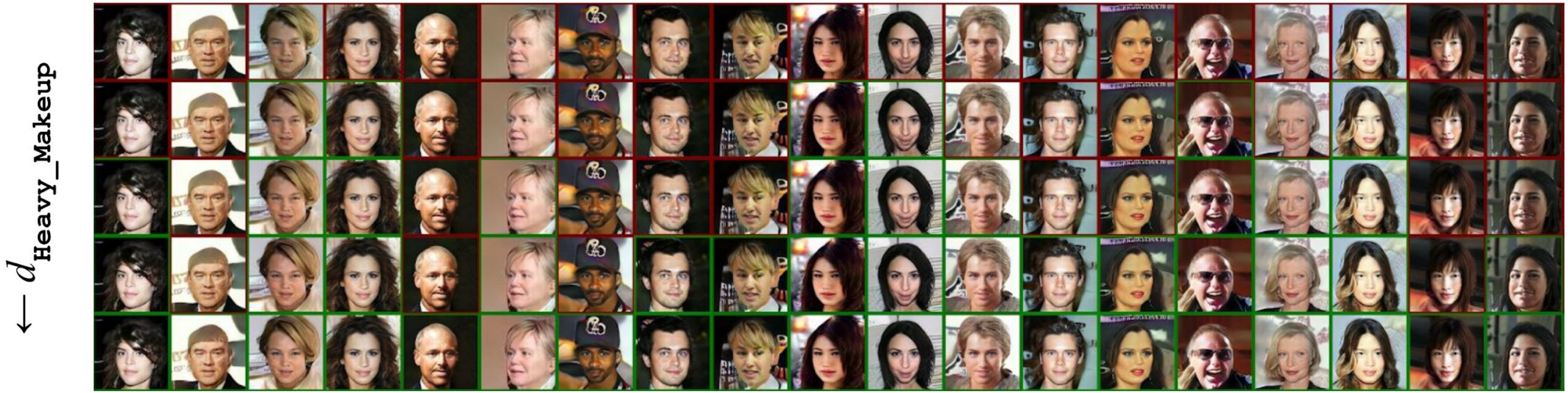
Sensitivity Results

$$S_f(d) = \mathbb{E}_{z \sim p(z)}[f(G(z+d)) - f(G(z))]$$





Heavy Makeup



Young

$\leftarrow d_{\texttt{Young}}$

Directional Sensitivity

$$S_y^{0\to 1}(d) = \mathbb{E}_{z\sim p(z)|y(G(z))=0}\ \mathbb{I}[y(G(z+d))! = y(G(z))]$$

from "not smiling" to "smiling"

$$S_y^{1\to 0}(d) = \mathbb{E}_{z\sim p(z)|y(G(z))=1}\ \mathbb{I}[y(G(z+d))! = y(G(z))]$$

from "smiling" to "not smiling"

$$S_f(d) = \mathbb{E}_{z\sim p(z)}[f(G(z+d)) - f(G(z))]$$

Sensitivity Results

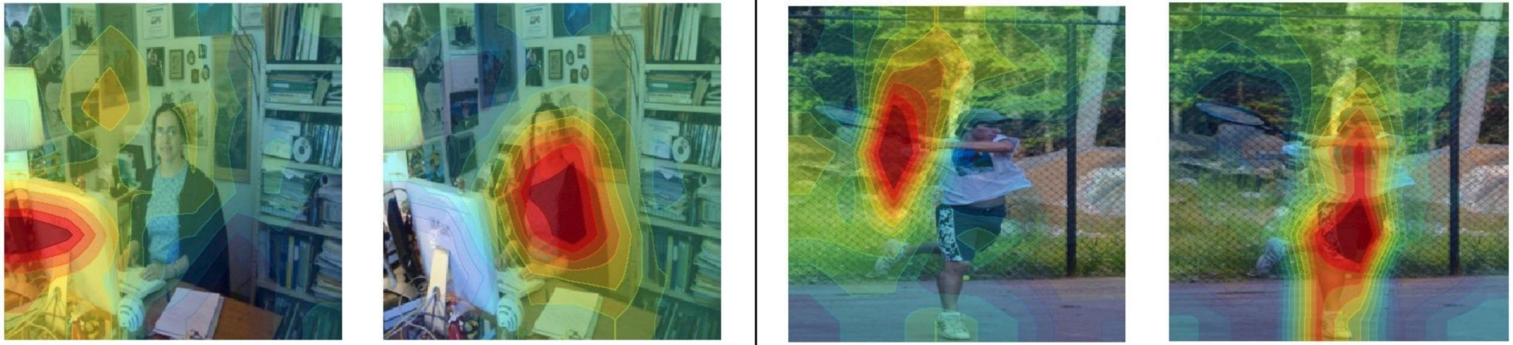
CelebA attribute defining d_a	$S_y^{1\to 0}(d_a)$	$S_y^{0\to 1}(d_a)$
Young	**7.0%**	2.6%
5_o_Clock_Shadow	**11.8%**	2.2%
Goatee	**12.4%**	0.9%
No_Beard	0.8%	**11.8%**
Heavy_Makeup	1.6%	**12.4%**
Wearing_Lipstick	1.7%	**16.3%**

Outline

- Counterfactual Fairness
- Counterfactual Face Attribution
- Gender Equalized Image Captioning
- Adversarial Removal of Gender Features

The Equalizer Model

Wrong | Right for the Right Reasons | Right for the Wrong Reasons | Right for the Right Reasons



Baseline:
*A **man** sitting at a desk with a laptop computer.*

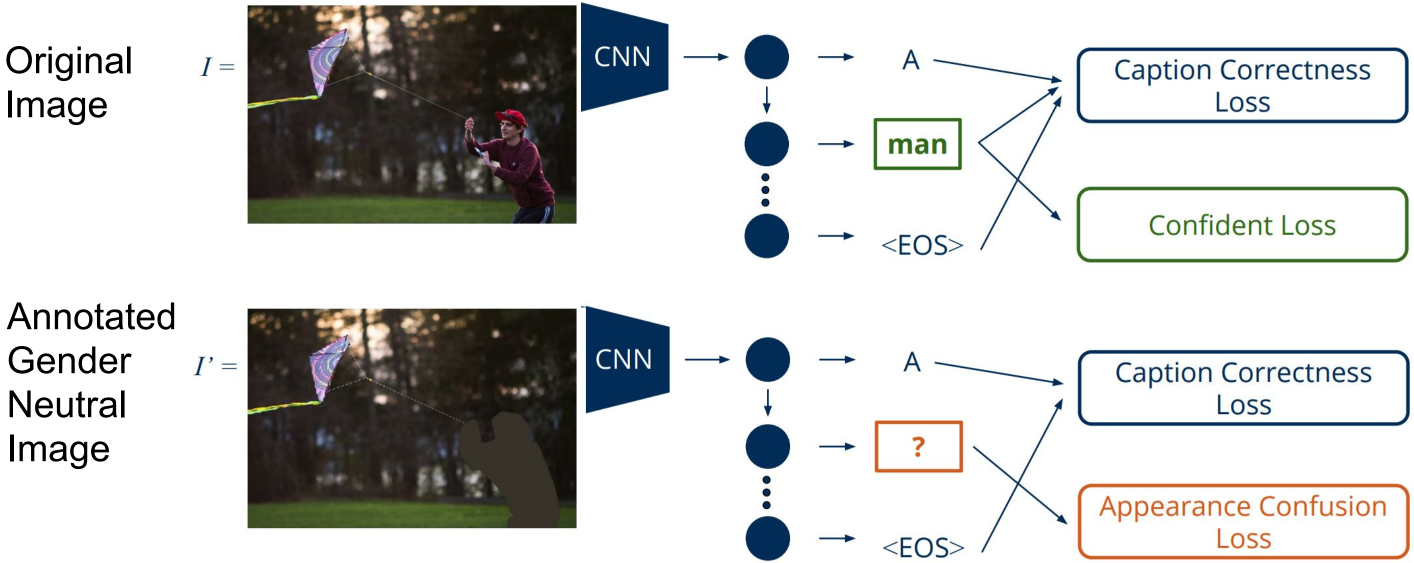
Our Model:
*A **woman** sitting in front of a laptop computer.*

Baseline:
*A **man** holding a tennis racquet on a tennis court.*

Our Model:
*A **man** holding a tennis racquet on a tennis court.*

[Burns et al, 2019]

The Basic Idea



Burns et al, 2019

The Equalizer Model

$$\mathcal{L} = \alpha\mathcal{L}^{CE} + \beta\mathcal{L}^{AC} + \mu\mathcal{L}^{Con}$$

Cross Entropy Loss

$$\mathcal{L}^{CE} = -\frac{1}{N}\sum_{n=0}^{N}\sum_{t=0}^{T}\log(p(w_t|w_{0:t-1}, I))$$

Appearance Confusing Loss on the gender neutral image

Confidence Loss on the original image

Burns et al, 2019

Appearance Confusing Objective

$$\mathcal{L}^{AC} = \frac{1}{N}\sum_{n=0}^{N}\sum_{t=0}^{T}\mathbb{1}(w_t \in \mathcal{G}_w \cup \mathcal{G}_m)\mathcal{C}(\tilde{w}_t, I')$$

$$\mathcal{C}(\tilde{w}_t, I') = |\sum_{g_w \in \mathcal{G}_w} p(\tilde{w}_t = g_w | w_{0:t-1}, I') - \sum_{g_m \in \mathcal{G}_m} p(\tilde{w}_t = g_m | w_{0:t-1}, I')|$$

Push Toward Extremes

$$p(\tilde{w}_t = g_w | w_{0:t-1}, I') \longleftrightarrow p(\tilde{w}_t = g_m | w_{0:t-1}, I')$$

$I' =$

$\mathcal{G}_w$ - set of words for woman

$\mathcal{G}_m$ - set of words for man

Burns et al, 2019

Confidence Objective

$$\mathcal{L}^{Con} = \frac{1}{N} \sum_{n=0}^{N} \sum_{t=0}^{T} (\mathbb{1}(w_t \in \mathcal{G}_w)\mathcal{F}^W(\tilde{w}_t, I) + \mathbb{1}(w_t \in \mathcal{G}_m)\mathcal{F}^M(\tilde{w}_t, I))$$

$$\mathcal{F}^W(\tilde{w}_t, I) = \frac{\sum_{g_m \in \mathcal{G}_m} p(\tilde{w}_t = g_m | w_{0:t-1}, I)}{(\sum_{g_w \in \mathcal{G}_w} p(\tilde{w}_t = g_w | w_{0:t-1}, I)) + \epsilon}$$

$I =$ 

$\mathcal{G}_w$ - set of words for woman
$\mathcal{G}_m$ - set of words for man

[Burns et al, 2019](#)

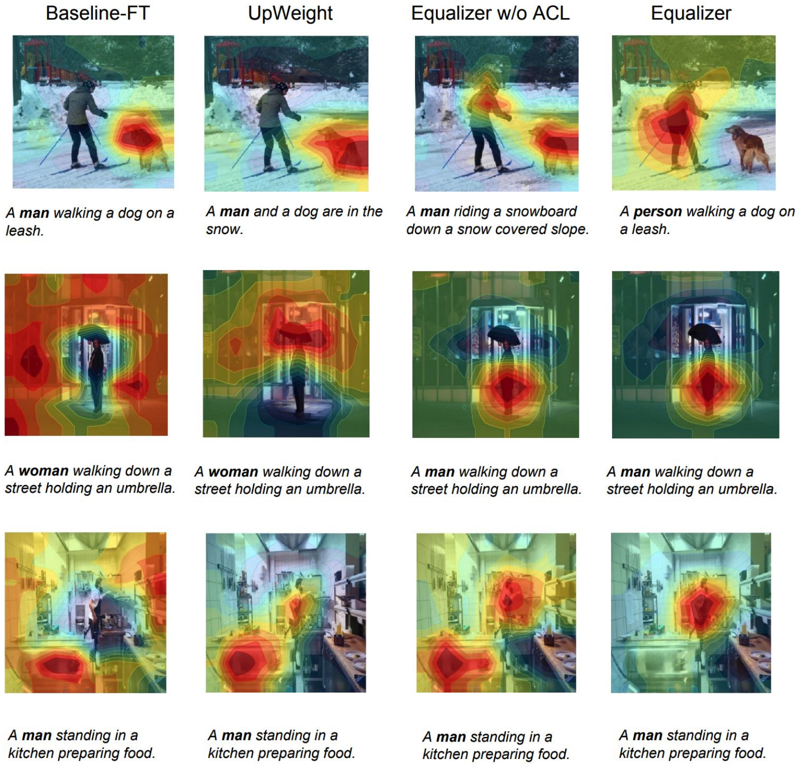
Results

Model	MSCOCO-Bias Error	MSCOCO-Bias Ratio Δ	MSCOCO-Balanced Error	MSCOCO-Balanced Ratio Δ
Baseline-FT	12.83	0.15	19.30	0.51
Balanced	12.85	0.14	18.30	0.47
UpWeight	13.56	0.08	16.30	0.35
Equalizer w/o ACL	7.57	0.04	10.10	0.26
Equalizer w/o Conf	9.62	0.09	13.90	0.40
Equalizer	**7.02**	**-0.03**	**8.10**	**0.13**

Baseline-FT - basic LSTM attention model (Xu et al, 2015)
Balanced - resampled dataset to have balanced gender ratio
UpWeight - reweighting
Δ - change to the gender ratio compared to the dataset

Burns et al, 2019

Results


Baseline-FT
UpWeight
Equalizer w/o ACL
Equalizer
A man walking a dog on a leash.
A man and a dog are in the snow.
A man riding a snowboard down a snow covered slope.
A person walking a dog on a leash.
A woman walking down a street holding an umbrella.
A woman walking down a street holding an umbrella.
A man walking down a street holding an umbrella.
A man walking down a street holding an umbrella.
A man standing in a kitchen preparing food.
A man standing in a kitchen preparing food.
A man standing in a kitchen preparing food.
A man standing in a kitchen preparing food.

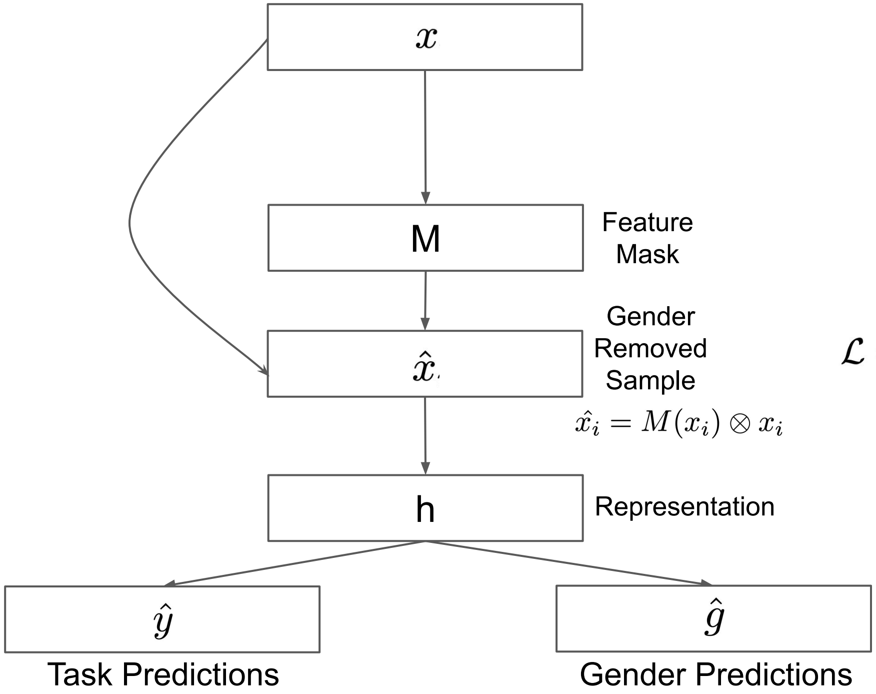
Outline

- Counterfactual Fairness
- Counterfactual Face Attribution
- Gender Equalized Image Captioning
- Adversarial Removal of Gender Features

Adversarial Removal of Gender Features

Wang et al, 2019

Model Architecture



$$\mathcal{L} = \sum_i \beta|x_i - \hat{x}_i| + \mathcal{L}_p(pred(h(\hat{x}_i)), y_i) - \lambda \mathcal{L}_c(c(h(\hat{x}_i)), g_i)$$

Reconstruction Loss

Task Loss

Adversarial Loss

Wang et al, 2019

Evaluate Sensitive Information Leakage

- Train an attacker f(y) that reverse engineer the gender information

DATASET LEAKAGE @ F1

knife
table
spoon
vase
cabinet
window

$\longrightarrow$ Randomly Perturb True Labels $r(Y, F1)$ $\longrightarrow$

knife
tv-set
spoon
vase
cabinet
hammer

$\longrightarrow$ Predict Gender from Labels $f(Y)$

$$\lambda_D(a) = \frac{1}{|\mathcal{D}|} \sum\nolimits_{(Y_i, g_i) \in \mathcal{D}} \mathbb{1}[f(r(Y_i, a)) == g_i]$$

MODEL LEAKAGE @ F1

$\longrightarrow$ Predict Labels from Image $p(X)$ $\longrightarrow$

knife
spoon
fork
vase
hair drier
oven

$\longrightarrow$ Predict Gender from Predicted Labels $f(\hat{Y})$

$$\lambda_M(a) = \frac{1}{|\mathcal{D}|} \sum\nolimits_{(\hat{Y}_i, g_i) \in \mathcal{D}} \mathbb{1}[f(r(\hat{Y}_i, a) == g_i)]$$

Data Resampling

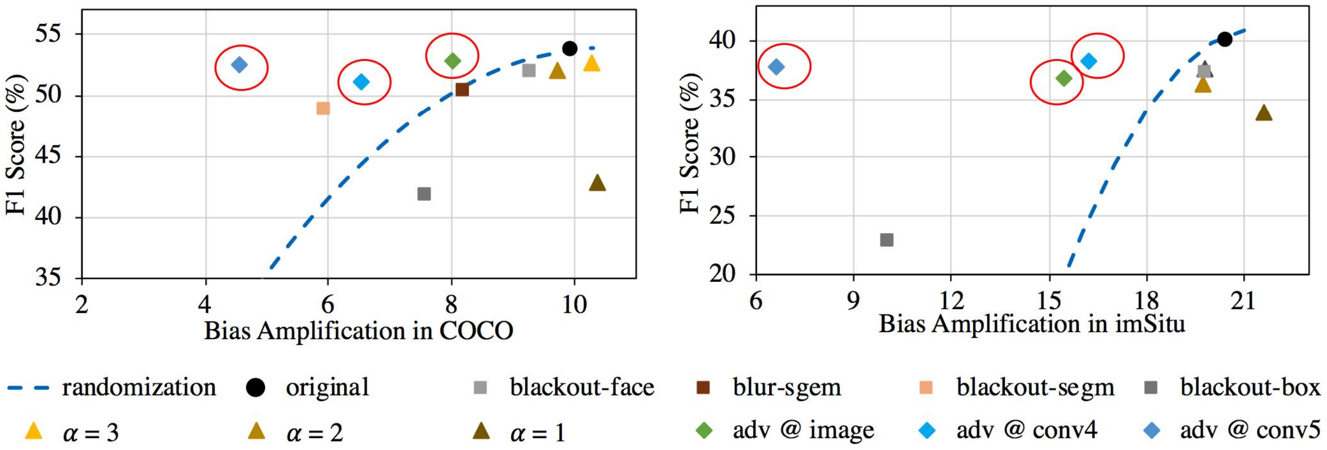
$$\forall y : 1/\alpha < \#(m, y)/\#(w, y) < \alpha$$

Bias Amplification

$$\Delta = \lambda_M(a) - \lambda_D(a)$$

Wang et al, 2019

Accuracy and Bias Results

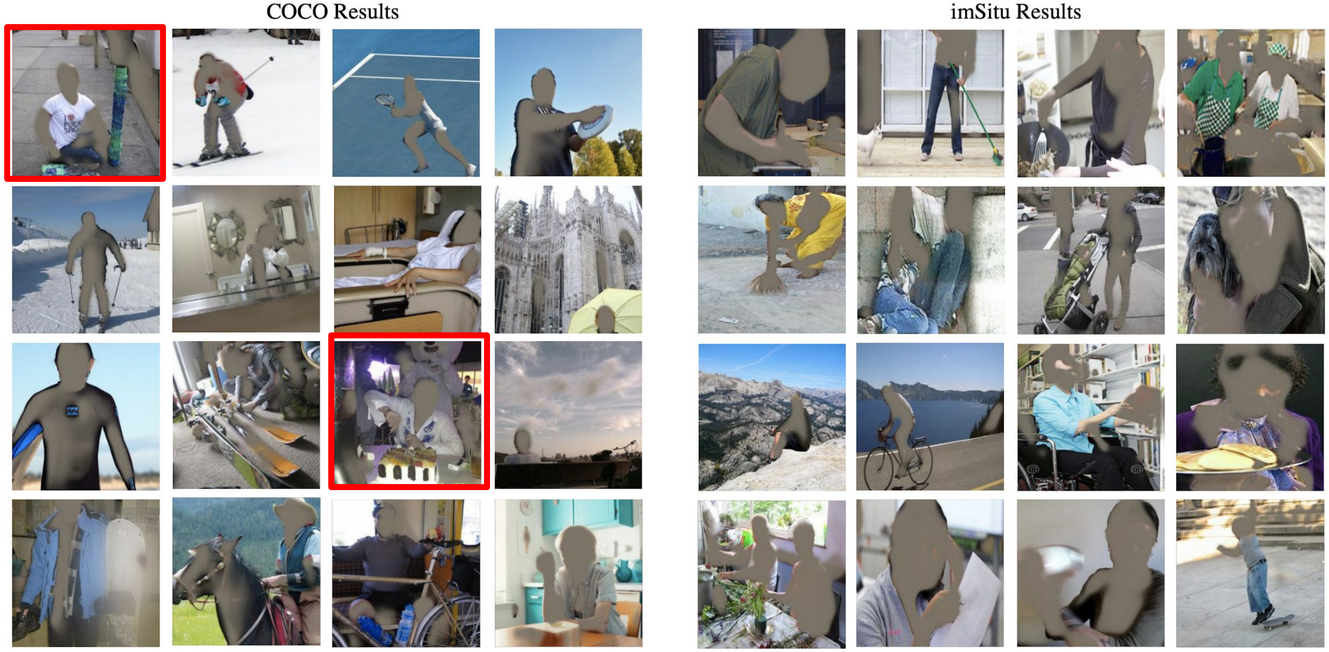


original - no debiasing mask
random - adding random noise

blackout-face - black out using a face detector
blur-sgem - black out using ground truth segmentation
blackout-box - blackout using bounding boxes

[Wang et al, 2019]

Qualitative Results



Wang et al, 2019

Summary

- Fair Machine Learning
  - Prevents ML models from biasing toward specific groups when allocating favorable outcomes
- Group Treatments of Fairness
  - Demographic Parity
  - Equalized Odds/Opportunity
- Individual Treatments of Fairness
  - Fairness Through Awareness Individual Fairness
  - Individual Fairness
  - Counterfactual Fairness
- Fair ML Techniques
  - Pre-processing Methods: Resampling, Reweighting, Optimized-preprocessing
  - In-processing Methods: Regularization, Adversarial Learning
  - Post-processing Methods: Learning to Defer

Summary

- Fair NLP Methods
  - Debiasing Word Embeddings
  - Data Augmentation
    - Gender Swapping
  - Fair Representation for Pre-trained Encoders

- Fair Visual Representations
  - Counterfactual Face Attribution
  - Gender Equalized Image Captioning
  - Adversarial Removal of Gender Features

Reading Assignments

- Kusner, Matt J., Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness, NeurIPS 2017
- Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints, EMNLP 2017
- Yin, Xi, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data, CVPR 2019
- Singh, Ashudeep, and Thorsten Joachims. Fairness of exposure in rankings, KDD 2017
- Buolamwini, Joy, and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification, FAccT 2018

Next Lecture

Mit-term Project Presentations