Optimization methods for imposing Fairness in Computer Vision Models

Utah Center for Data Science

September, 18th 2020

Vishnu Lokhande
UW-Madison
lokhande@cs.wisc.edu

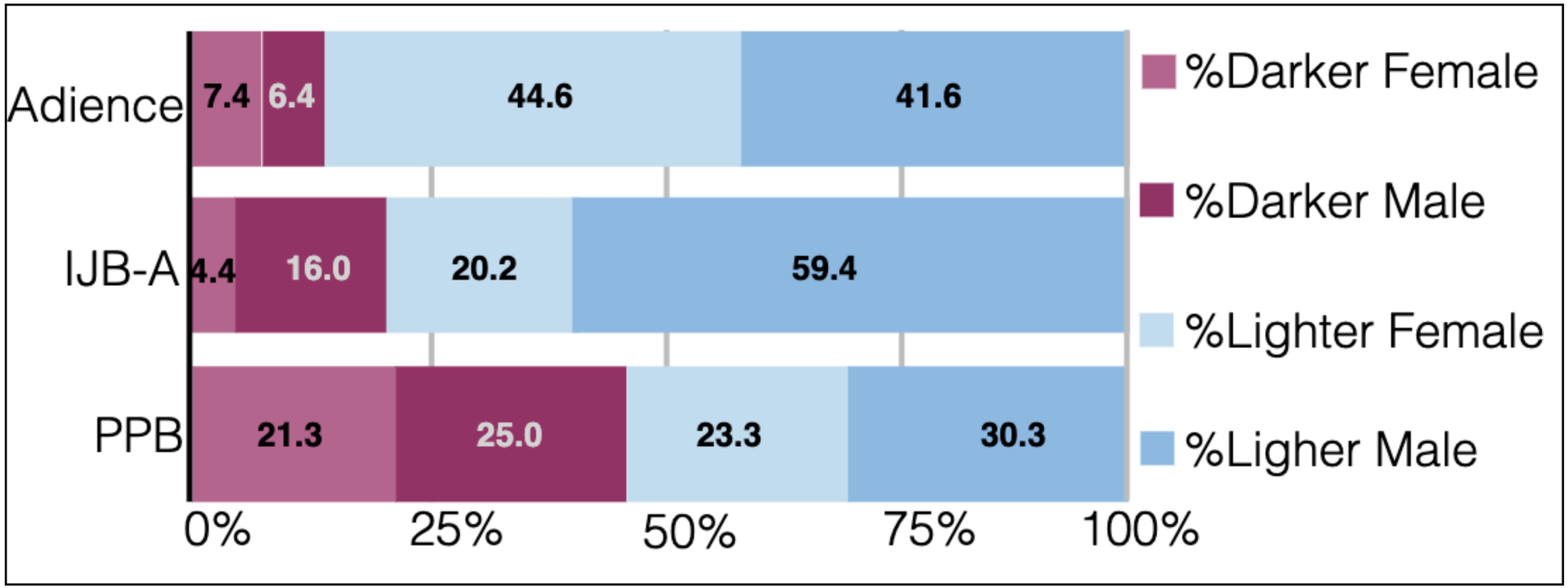
Wrongfully Accused by an Algorithm

In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

* New York Times, June 24th 2020

Dataset Bias

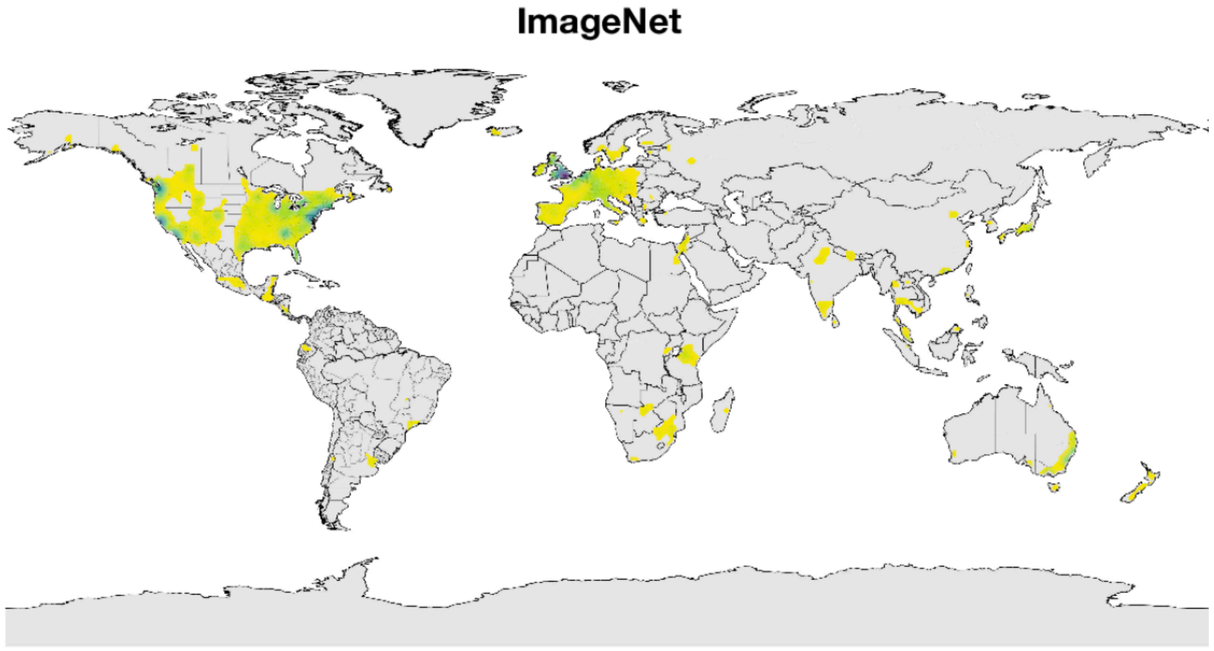
Race Diversity

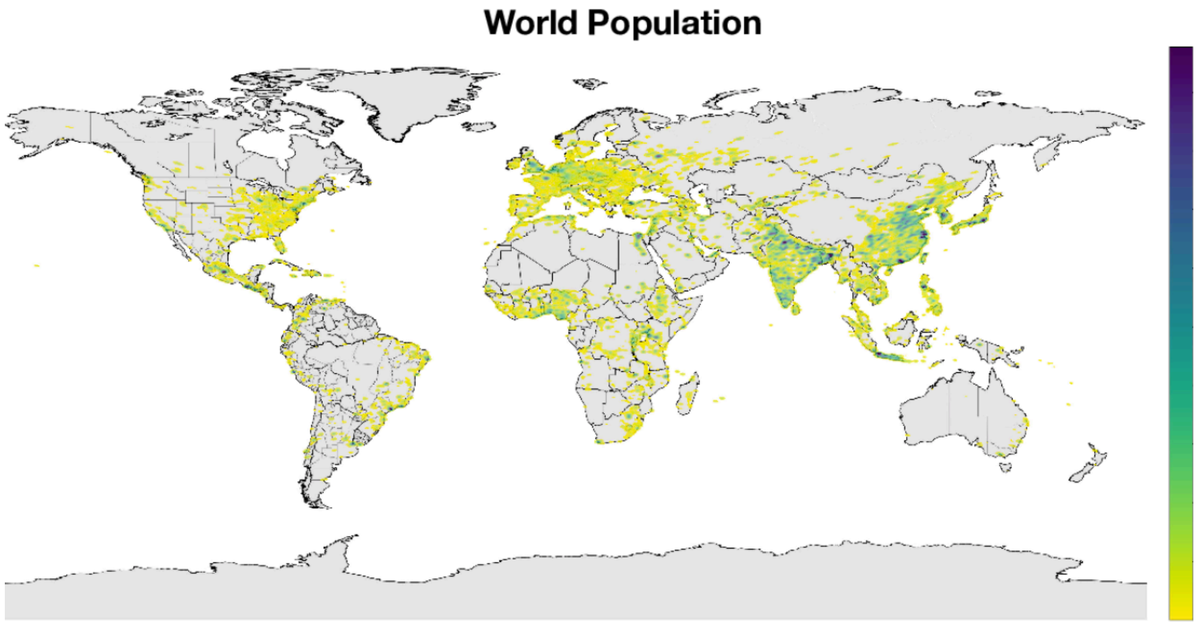


Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." *Conference on fairness, accountability and transparency*. 2018.

Dataset Bias

Geographic Diversity

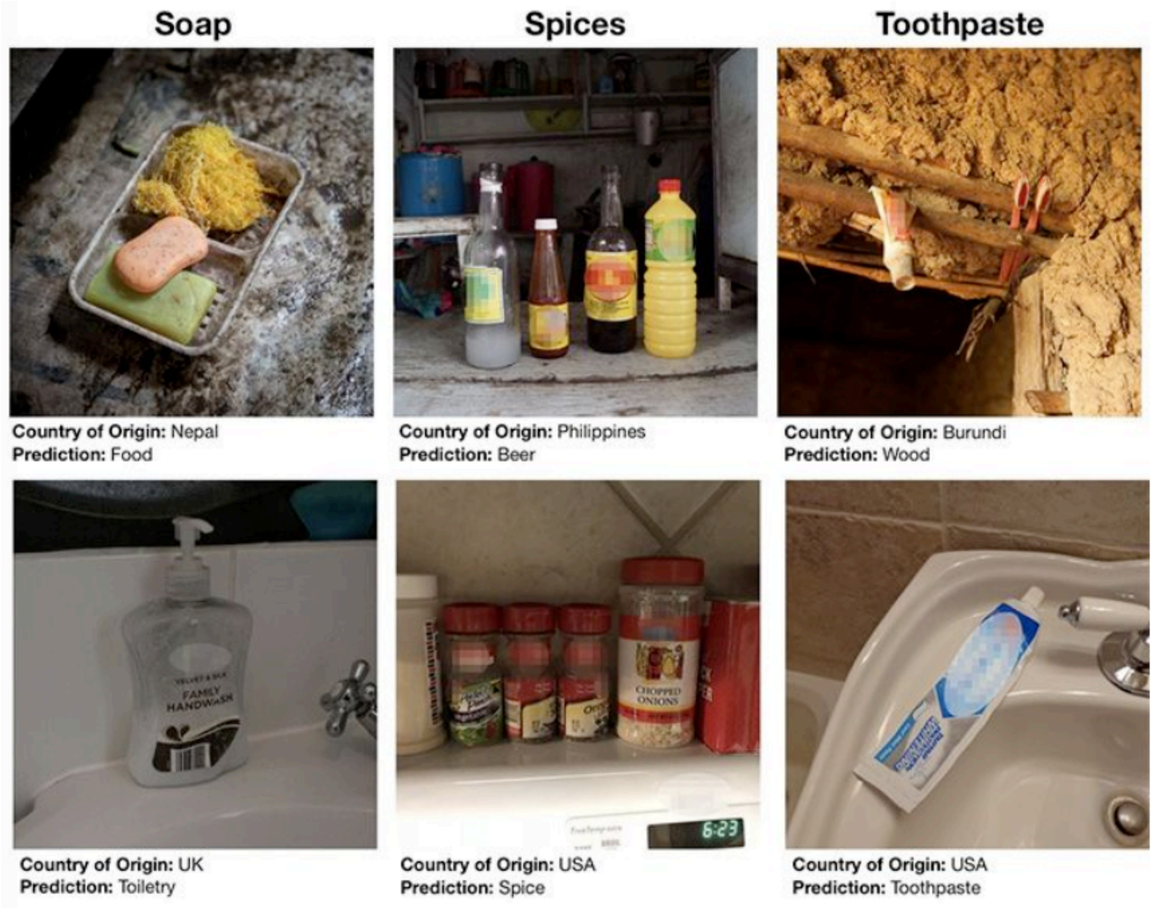




de Vries, Terrance, et al. "Does object recognition work for everyone?." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019.

Dataset Bias

Geographic Diversity



de Vries, Terrance, et al. "Does object recognition work for everyone?." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019.

General Strategies

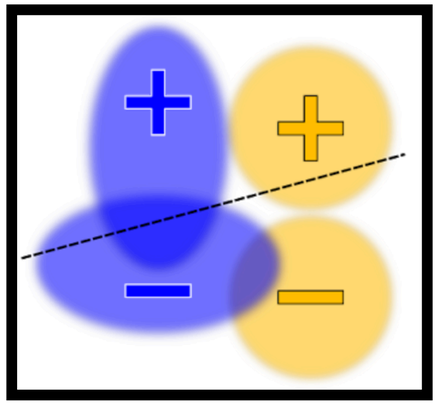
Pre-Processing

In-Processing

Algorithm 2 FairALM: DeepNet Classifier

Notations: Dual step size η, Primal step size τ
Input: Training Set D
Initializations: $\lambda_0 = 0$, w_0
for $t = 0, 1, 2, ..., T$ **do**
  Sample $z \sim D$
  Pick $v_t \in \partial\Big(\hat{e}_{h_w}(z) + (\lambda_t + \eta)\hat{\mu}^{s_0}_{h_w}(z) - (\lambda_t - \eta)\hat{\mu}^{s_1}_{h_w}(z)\Big)$
  (Primal) $w_t \leftarrow w_{t-1} - \tau v_t$
  (Dual) $\lambda_{t+1} \leftarrow \lambda_t + \eta(\hat{\mu}^{s_0}_{h_{w_t}}(z) - \hat{\mu}^{s_1}_{h_{w_t}}(z))$
end for
Output: w_T

Post-Processing

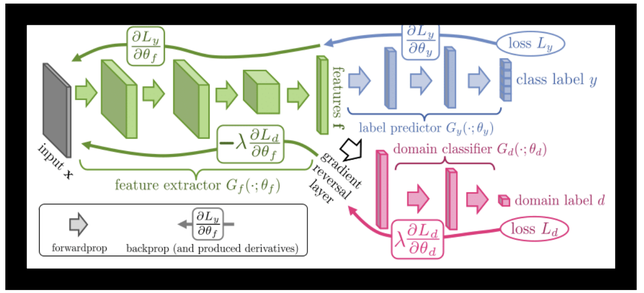


Quadrianto, Novi, Viktoriia Sharmanska, and Oliver Thomas. "Discovering fair representations in the data domain." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

Cotter, Andrew, et al. "Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals." Journal of Machine Learning Research 20.172 (2019): 1-59.

General Strategies

Adversarial Methods



FairALM

$$\min_{h \in \mathcal{H}} \quad e_h$$

$$\mu_h^{s_0} = \mu_h^{s_1}$$

Simpler Alternative

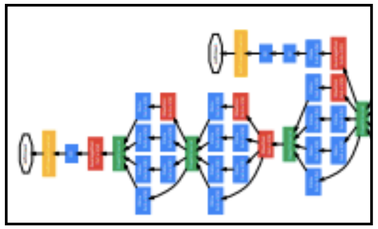
Ganin, Yaroslav, et al. "Domain-adversarial training of neural networks." *The Journal of Machine Learning Research* 17.1 (2016): 2096-2030.

- **Notations**
- The Lagrangian Formulation
- FairALM
    - Linear Classifiers
    - Deep Networks
- Experimental Results

The Expected Loss function

(x, y) : (Features, Label) → ( , 1)

s : Sensitive Attribute → Female

h : Classifier → 

$$\mathop{\mathbb{E}}_{x,y,s} \mathscr{L}(h; (x, y))$$

Fairness as Equality of Conditional Means

$$\mu_h^{s_0} = \mu_h^{s_1}$$

Demographic Parity	$\mu_h^s := \quad e_h \mid (s)$
Equality of Opportunity	$\mu_h^s := \quad e_h \mid (s, y)$
Predictive Parity	$\mu_h^s := \quad e_h \mid (s, \hat{y})$

The Constrained Optimization Problem

$$\min_{h \in \mathcal{H}} \quad e_h$$

$$\mu_h^{s_0} = \mu_h^{s_1}$$

- **Notations**
- **The Lagrangian Formulation**
- FairALM
  - Linear Classifiers
  - Deep Networks
- Experimental Results

The Lagrangian Optimization Problem

$$L(h, \lambda) = e_h + \lambda(\mu_h^{s_0} - \mu_h^{s_1})$$

$$\max_{\lambda \in R} \quad \underbrace{\min_{h \in \mathcal{H}} L(h, \lambda)}$$

Non-Smooth

Using Dual Proximal Functions

$$L_{\lambda_T}(h, \lambda) = e_h + \lambda(\mu_h^{s_0} - \mu_h^{s_1}) - \frac{1}{2\eta}(\lambda - \lambda_T)^2$$

- **Notations**
- **The Lagrangian Formulation**
- **FairALM**
  - **Linear Classifiers**
  - **Deep Networks**
- **Experimental Results**

FairALM: Linear Ensemble Classifiers

$$\mathscr{H} = \{h_1, h_2, h_3, \ldots, h_N\}$$
$$\downarrow \quad \downarrow \quad \downarrow \qquad \downarrow$$
$$e_{h_1}, e_{h_2}, e_{h_3}, \ldots, e_{h_N}$$

Agarwal, Alekh, et al. "A reductions approach to fair classification." arXiv preprint arXiv:1803.02453 (2018).

FairALM: Linear Ensemble Classifiers

$$L_{\lambda_T}(h, \lambda) = e_h + \lambda(\mu_h^{s_0} - \mu_h^{s_1}) - \frac{1}{2\eta}(\lambda - \lambda_T)^2$$

$$q \in \Delta^N$$

$$L_{\lambda_T}(q, \lambda) = \sum_i q_i \, e_{h_i} + \sum_i q_i \, \lambda(\mu_{h_i}^{s_0} - \mu_{h_i}^{s_1}) - \frac{1}{2\eta}(\lambda - \lambda_T)^2$$

FairALM: Linear Ensemble Classifier

$$L_{\lambda_T}(q,\lambda) = \sum_i q_i\, e_{h_i} + \sum_i q_i\, \lambda(\mu_{h_i}^{s_0} - \mu_{h_i}^{s_1}) - \frac{1}{2\eta}(\lambda - \lambda_T)^2$$

$$\max_{\lambda \in R} \min_{q \in \Delta^N} L_{\lambda_T}(q,\lambda)$$

Linear Program in q	$\text{argmin}_i \quad L_{\lambda_t}(q_i, \lambda_t)$
Maximize the Cumulative Reward	$\lambda_{t+1} \leftarrow \lambda_t + \frac{\eta}{t}(\mu_{h_t}^{s_0} - \mu_{h_t}^{s_1})$

Convergence Analysis

After T update steps $\Longrightarrow$ ν – approximate saddle point

$$L_{\lambda_T}\left(\frac{1}{T}\sum q_t, \quad \frac{1}{T}\sum \lambda_t\right) \leq \min_q L_{\lambda_T}\left(q, \quad \frac{1}{T}\sum \lambda_t\right) + \nu$$

$$L_{\lambda_T}\left(\frac{1}{T}\sum q_t, \quad \frac{1}{T}\sum \lambda_t\right) \geq \max_\lambda L_{\lambda_T}\left(\frac{1}{T}\sum q_t\ , \lambda\right) - \nu$$

$$\nu = \mathcal{O}\left(\frac{\log^2 T}{T}\right)$$

vs. $\mathcal{O}\left(\frac{\sqrt{T}}{T}\right)$ [Agarwal et al.]

- Notations
- The Lagrangian Formulation
- FairALM
    - Linear Classifiers
    - Deep Networks
- Experimental Results

Adjustments for Deep Networks - 1

- Replace non-differentiable indicator function to a smooth surrogate function like logistic function.

- Replace Error and Conditional Means with Empirical Estimates

$$e_h \rightarrow \hat{e}_{h_w}$$

$$\mu_h \rightarrow \hat{\mu}^s_{h_w}$$

Adjustments for Deep Networks - 2

$$\max_{\lambda} \min_{w} \left(\hat{e}_{h_w} + \lambda(\hat{\mu}^{s_0}_{h_w} - \hat{\mu}^{s_1}_{h_w}) - \frac{1}{2\eta}(\lambda - \lambda_t)^2 \right)$$

$$\leq \min_{w} \max_{\lambda} \left(\hat{e}_{h_w} + \lambda(\hat{\mu}^{s_0}_{h_w} - \hat{\mu}^{s_1}_{h_w}) - \frac{1}{2\eta}(\lambda - \lambda_t)^2 \right)$$

$$\lambda \leftarrow \lambda_t + \eta(\hat{\mu}^{s_0}_{h_w} - \hat{\mu}^{s_1}_{h_w})$$

Dual Update

$$\leq \min_{w} \left(\hat{e}_{h_w} + (\lambda_t + \eta)\hat{\mu}^{s_0}_{h_w} - (\lambda_t - \eta)\hat{\mu}^{s_1}_{h_w} \right)$$

FairALM Objective

FairALM: DeepNet Classifier

1: $\lambda_0 = 0, \quad w_0 = 0$

2: **FOR** $t = 0,1,2,\ldots,T$ **DO**

3: **Sample** $z \sim$ **Data**

4: **Pick** $v_t \in \partial\left(\hat{e}_{h_w} + (\lambda_t + \eta)\hat{\mu}^{s_0}_{h_w} - (\lambda_t - \eta)\hat{\mu}^{s_1}_{h_w}\right)$

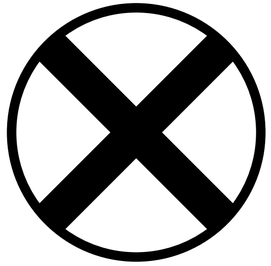
5: $w_t \leftarrow w_{t-1} - \tau v_t$ Primal Update

6: $\lambda \leftarrow \lambda_t + \eta(\hat{\mu}^{s_0}_{h_w} - \hat{\mu}^{s_1}_{h_w})$ Dual Update

7: **END**

Unconstrained

1: $w_0 = 0$
2: **FOR** $t = 0,1,2,\ldots,T$ **DO**
3: **Sample** $z \sim$ **Data**
4: **Pick** $v_t \in \partial\left(\hat{e}_{h_w}\right)$
5: $w_t \leftarrow w_{t-1} - \tau v_t$
6:
6: **END**



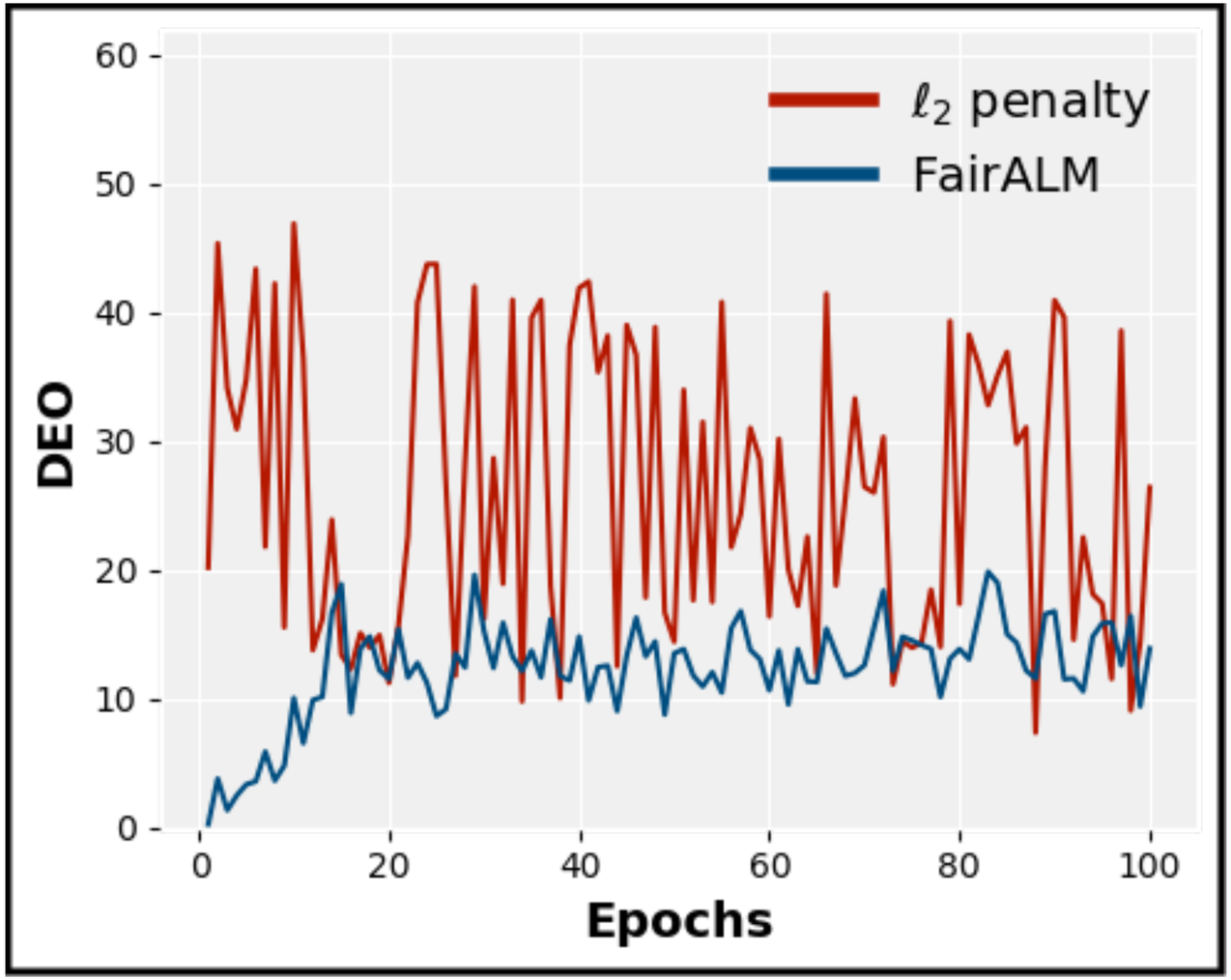
FairALM

1: $\lambda_0 = 0, \quad w_0 = 0$
2: **FOR** $t = 0,1,2,\ldots,T$ **DO**
3: **Sample** $z \sim$ **Data**
4: **Pick** $v_t \in \partial\left(\hat{e}_{h_w} + (\lambda_t + \eta)\hat{\mu}^{s_0}_{h_w} - (\lambda_t - \eta)\hat{\mu}^{s_1}_{h_w}\right)$
5: $w_t \leftarrow w_{t-1} - \tau v_t$
6: $\lambda \leftarrow \lambda_t + \eta(\hat{\mu}^{s_0}_{h_w} - \hat{\mu}^{s_1}_{h_w})$
7: **END**

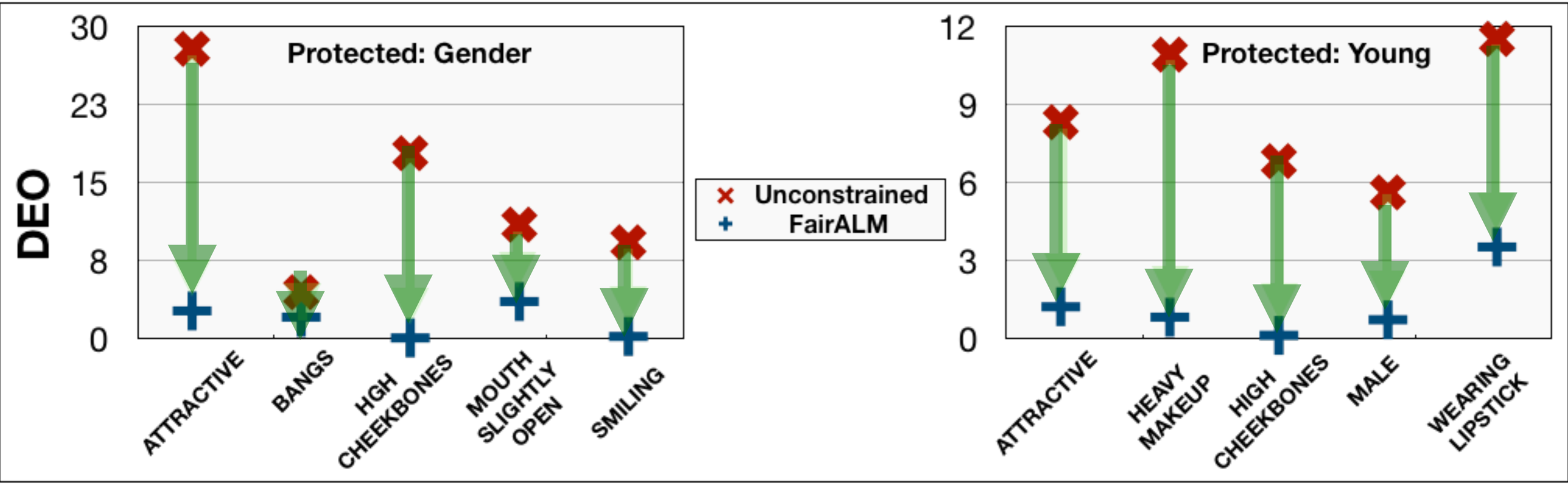
- Notations
- The Lagrangian Formulation
- FairALM
    - Linear Classifiers
    - Deep Networks
- Experimental Results

CELEBA DATASET

A STABLE TRAINING PROFILE

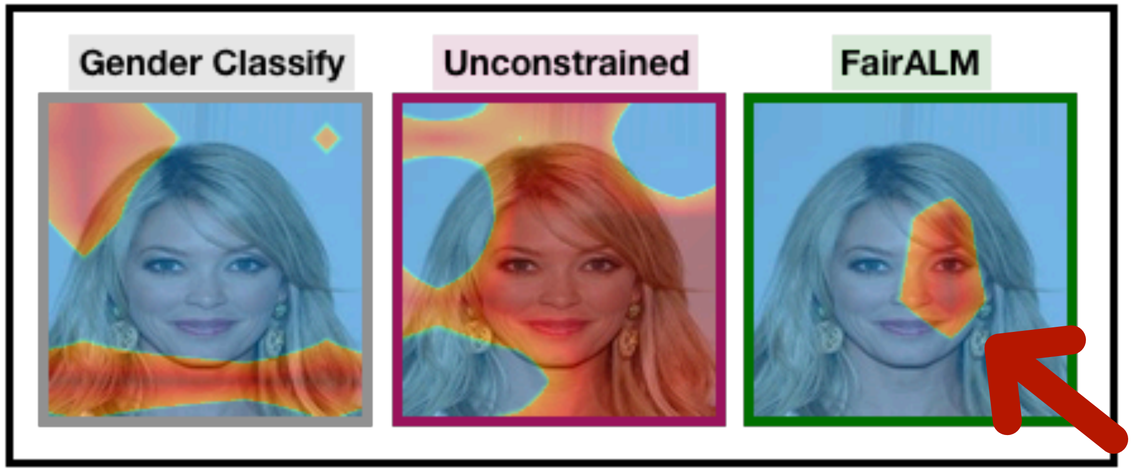


CELEBA DATASET



CELEBA DATASET

Interpretable Models



IMSITU DATASET

INTERPRETABILITY EXAMPLES - 1/7

Cooking (+)
Driving (-)

Unconstrained

Focus: **PERSON**
Feature leaking Gender

FairALM

Focus: **FOOD**
Feature concealing Gender

IMSITU DATASET

INTERPRETABILITY EXAMPLES - 2/7

Microwaving (+)
Pumping (-)

Unconstrained

Focus: **FACE**
Feature revealing Gender

FairALM

Focus: **MICROWAVE/UTENSILS**
Feature concealing Gender

IMSITU DATASET

INTERPRETABILITY EXAMPLES - 3/7

Cooking (+)
Driving (-)

Unconstrained

Focus: **PERSON**
Feature leaking Gender

FairALM

Focus: **FOOD**
Feature concealing Gender

IMSITU DATASET

INTERPRETABILITY EXAMPLES - 4/7

Driving (+)
Cooking (-)

Unconstrained

Focus: **HAIR**
Feature related to Gender

FairALM

Focus: **STEERING WHEEL**
Feature concealing Gender

IMSITU DATASET

INTERPRETABILITY EXAMPLES - 5/7

Shaving (+)
Moisturizing (-)

Unconstrained

Focus: **NAIL POLISH**
Feature related to Gender

FairALM

Focus: **RAZOR**
Feature concealing Gender

IMSITU DATASET

INTERPRETABILITY EXAMPLES - 6/7

Shaving (+)
Moisturizing (-)

Unconstrained

Focus: **FOREHEAD**
Inaccurate feature

FairALM

Focus: **RAZOR/SHAVING CREAM**
Feature concealing Gender

IMSITU DATASET

INTERPRETABILITY EXAMPLES - 7/7

Assembling (+)
Hanging (-)

Unconstrained

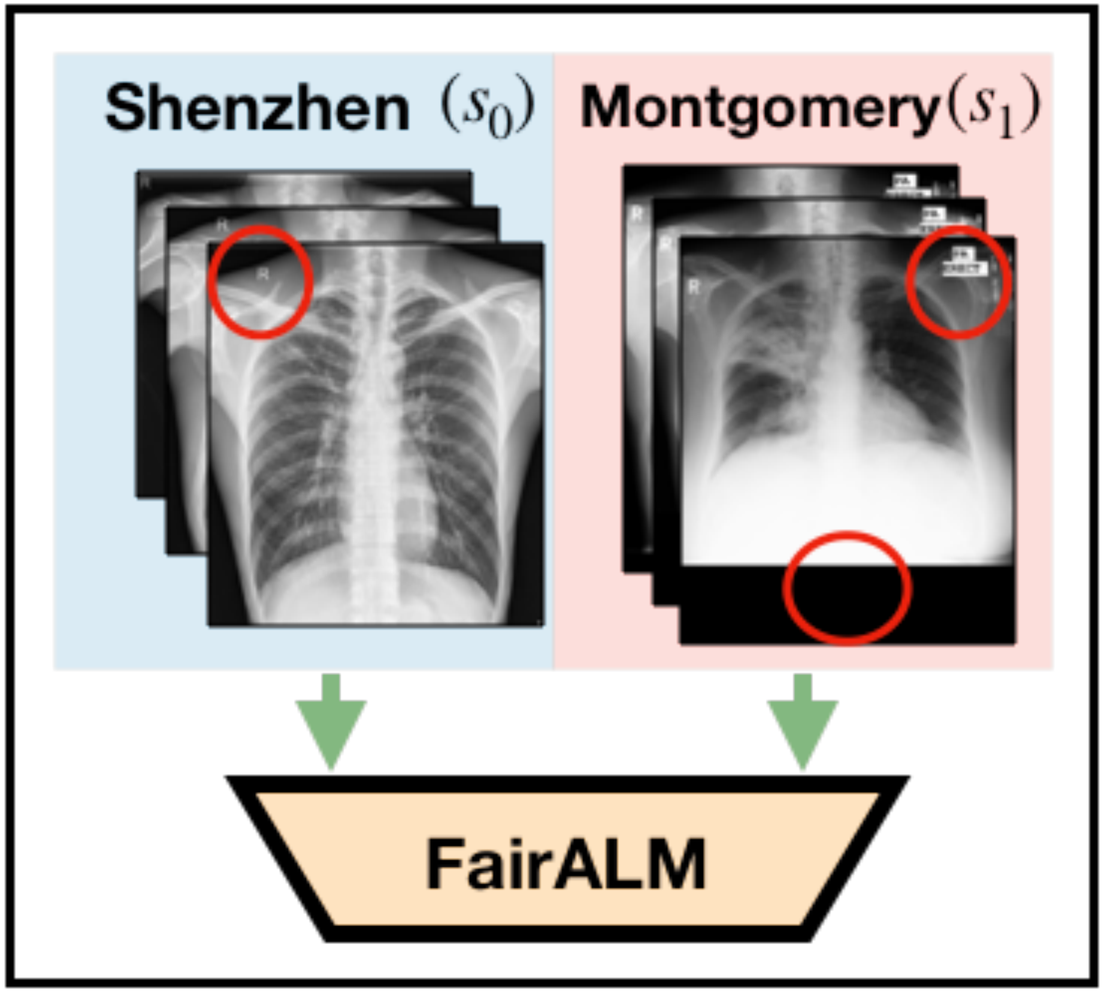
Focus: **HELMET**
Feature less aligned with the Target label Assembling

FairALM

Focus: **HELMET+TOOLS**
Feature more aligned with the Target label Assembling

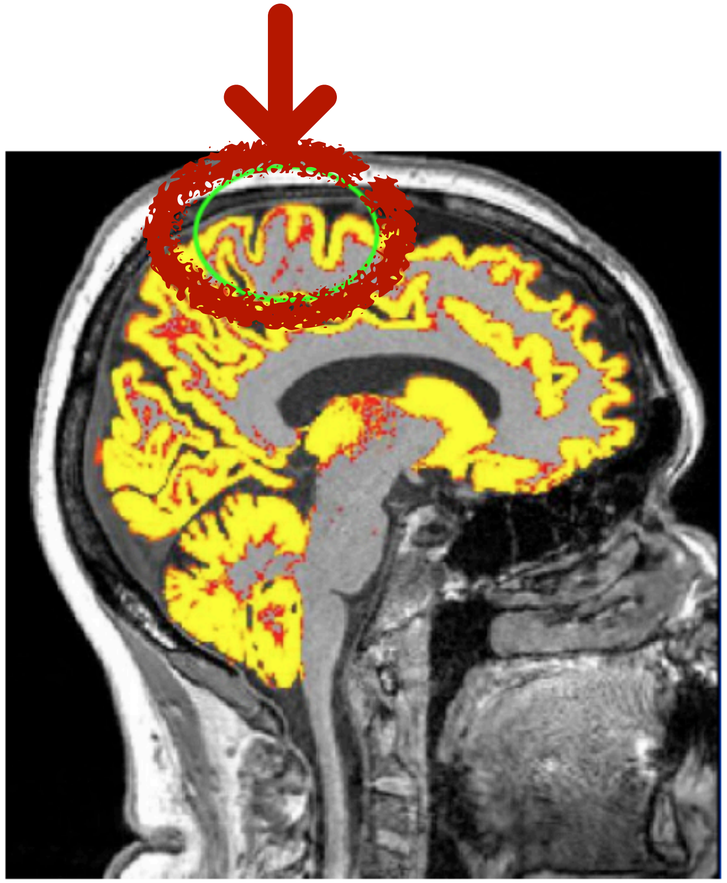
BEYOND FAIRNESS

MULTI-SITE POOLING - 1/2

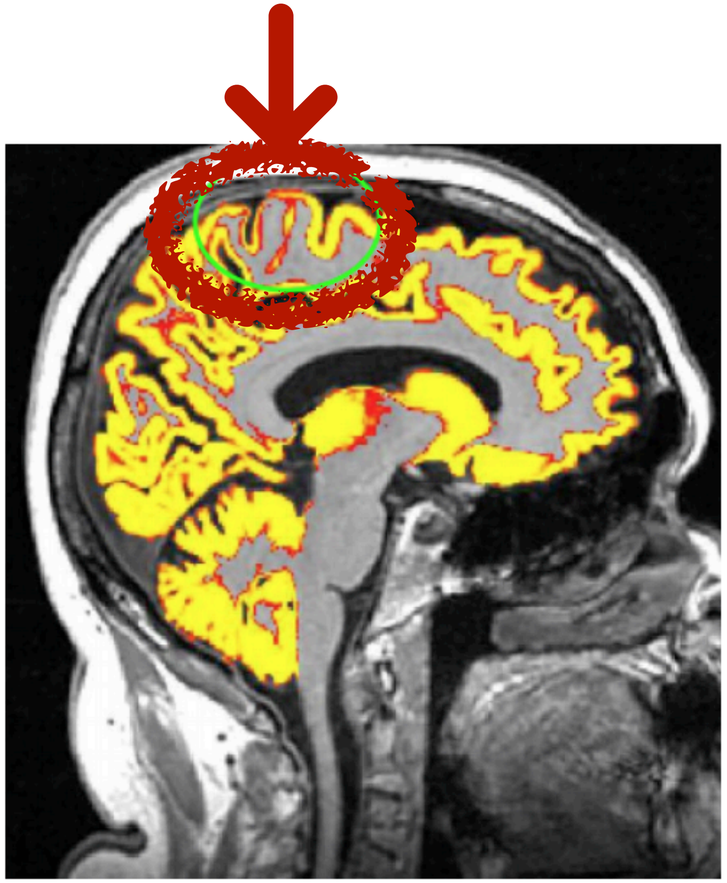


BEYOND FAIRNESS

MULTI-SITE POOLING - 2/2



Scanner: GE



Scanner: Siemens

Ite Missa Est