Mitigating Bias Using Model-Agnostic Data Attribution

Sander De Coninck, Sam Leroux, Pieter Simoens
Sander.DeConinck@UGent.be, UGent - imec

CVPR SEATTLE, WA JUNE 17-21, 2024
Workshop On Fair, Data-efficient, And Trusted Computer Vision

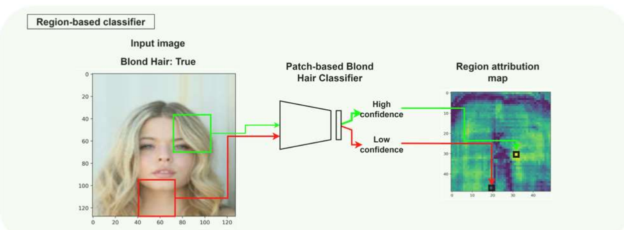
Introduction

Goal?: learning unbiased classifiers using on data where a confounding attribute biases the data.

How? By preventing models from overfitting on the confounders using targeted noise.

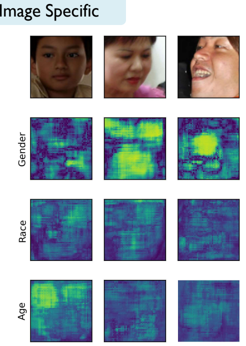
Challenge? Finding what image pixels contribute towards being able to classify an attribute and ensuring models don't utilize these

Model-Agnostic Data Attribution



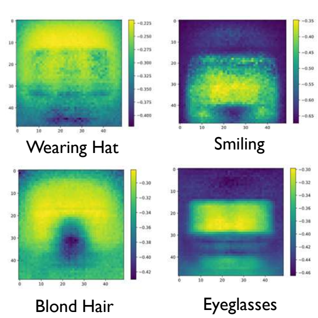
Attribution Visualizations

Image Specific



FairFace[1] attributes

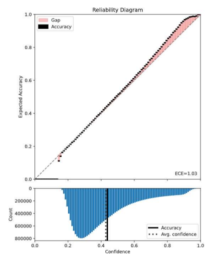
Averaged General



CelebA[2] attributes

Calibration

Training on random images patches ensures calibrated models



Reliability diagram for FairFace[1] race attribute

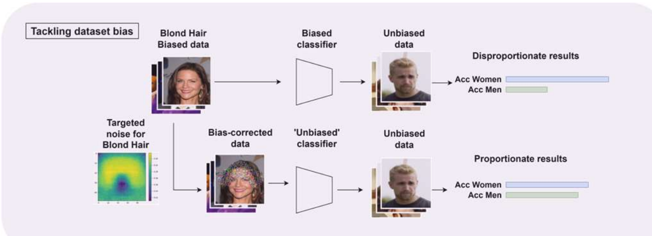
Sanity check

If we add noise to our attributed regions, does the accuracy decrease more than when we add noise randomly?

Attribute	Mask type	Accuracy Targeted	Accuracy Random	Δ
Blond Hair	General	0.78	0.85	-0.07
Eyeglasses	General	0.62	0.80	-0.18
Smiling	General	0.63	0.74	-0.11
Wearing Hat	General	0.66	0.87	-0.20
Blond Hair	Specific	0.80	0.85	-0.04
Eyeglasses	Specific	0.59	0.80	-0.22
Smiling	Specific	0.51	0.74	-0.22
Wearing Hat	Specific	0.69	0.87	-0.18

Yes, although there are significant differences between attributes.

Bias Mitigating Learning



Experiment Setup

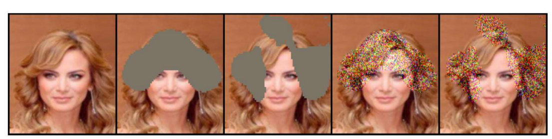
Task: perceived gender classification.

Data: biased subset of CelebA[2]

- 3000 Men: 0 with confounder
- 3000 Women: 2000 with confounder

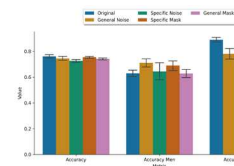
Goal: comparable accuracy across men and women when evaluating on balanced data

Adding noise based on attributions

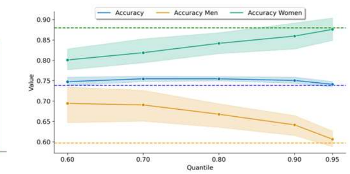


Experiment Results

Confounder: Blond Hair



Our technique decreases the accuracy gap between subgroups



The size of the region is another hyperparameter

Large scale experiment

Attribute	Noise	Type Quantile	Accuracy ↑ Original	Accuracy ↑ Balanced Δ	Accuracy ↑ Ours Δ	Accuracy Men ↑ Original	Accuracy Men ↑ Balanced Δ	Accuracy Men ↑ Ours Δ	Accuracy Women ↑ Original	Accuracy Women ↑ Balanced Δ	Accuracy Women ↑ Ours Δ	Gap ↓ Original	Gap ↓ Balanced	Gap ↓ Ours
Blond Hair	General Mask	0.60	0.74	+0.09	0.0	0.6	+0.2	+0.02	0.88	-0.03	-0.03	0.28	0.05	0.22
	General Noise	0.70	0.77	+0.06	-0.02	0.64	+0.16	+0.08	0.9	-0.05	-0.12	0.26	0.05	0.09
	Specific Mask	0.60	0.74	+0.09	+0.01	0.6	+0.2	+0.09	0.88	-0.03	-0.08	0.28	0.05	0.11
	Specific Noise	0.80	0.79	+0.04	-0.05	0.67	+0.13	+0.01	0.9	-0.05	-0.1	0.23	0.05	0.15
Eyeglasses	General Mask	0.60	0.71	+0.06	-0.02	0.61	+0.11	+0.03	0.82	+0.0	-0.07	0.21	0.10	0.12
	General Noise	0.60	0.72	+0.05	+0.01	0.61	+0.11	+0.05	0.84	-0.02	-0.04	0.23	0.10	0.14
	Specific Mask	0.60	0.71	+0.06	+0.03	0.61	+0.11	+0.13	0.82	+0.0	-0.08	0.21	0.10	0.05
	Specific Noise	0.80	0.72	+0.05	-0.01	0.61	+0.11	+0.1	0.84	-0.02	-0.13	0.23	0.10	0.09
Smiling	General Mask	0.80	0.81	+0.03	-0.04	0.71	+0.06	-0.02	0.9	+0.01	-0.05	0.19	0.13	0.16
	General Noise	0.60	0.85	-0.01	-0.05	0.79	-0.02	-0.04	0.91	-0.0	-0.07	0.12	0.13	0.09
	Specific Mask	0.95	0.81	+0.03	-0.02	0.71	+0.06	-0.02	0.9	+0.01	-0.01	0.19	0.13	0.20
	Specific Noise	0.60	0.85	-0.01	-0.05	0.79	-0.02	+0.01	0.91	-0.0	-0.1	0.12	0.13	0.05
Wearing Hat	General Mask	0.80	0.71	+0.08	+0.05	0.63	+0.11	+0.08	0.79	+0.05	+0.03	0.16	0.10	0.10
	General Noise	0.95	0.73	+0.06	-0.02	0.63	+0.11	+0.03	0.84	+0.0	-0.07	0.21	0.10	0.11
	Specific Mask	0.95	0.71	+0.08	0.0	0.63	+0.11	+0.01	0.79	+0.05	-0.01	0.16	0.10	0.15
	Specific Noise	0.95	0.73	+0.06	-0.03	0.63	+0.11	+0.04	0.84	+0.0	-0.1	0.21	0.10	0.08

- Our results approach balanced scenario
- Even on balanced data gap remains
- Noise scheme & quantile remain hyperparameters

Conclusion & Future Work

We introduced a novel way to **prevent bias** when training on **heavily unbalanced data** through **additive noise** on regions influencing the confounding attribute

We obtained these attributions in a model-agnostic way, by using the confidence of a **well-calibrated patch-based classifier**.

We believe our technique for attributing image data could see use in more areas of the fair, data-efficient and trusted research, notably in **privacy-preserving computer vision**

References

1. Karkkainen, Kimmo, and Jungseock Joo. "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation." *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2021.
2. Liu, Ziwei, et al. "Deep learning face attributes in the wild." *Proceedings of the IEEE international conference on computer vision*. 2015.