FairCal: Fairness Calibration For Face Verification

Removing bias through clustering and calibration

Tiago Salvador[1,3], Stephanie Cairns[1,3], Vikram Voleti[2,3], Noah Marshall[1,3], Adam Oberman[1,3]

[1] McGill University [2] Université de Montréal [3] Mila

Bias in AI

Overcoming Racial Bias In AI Systems And Startlingly Even In AI Self-Driving Cars

Racial bias in a medical algorithm favors white patients over sicker black patients

AI expert calls for end to UK use of 'racially biased' algorithms

AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators

Gender bias in AI: building fairer algorithms

Bias in AI: A problem recognized but still unresolved

Amazon, Apple, Google, IBM, and Microsoft worse at transcribing black people's voices than white people's with AI voice recognition, study finds

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

The Week in Tech: Algorithmic Bias Is Bad. Uncovering It Is Good.

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

Artificial Intelligence has a gender bias problem – just ask Siri

The Best Algorithms Struggle to Recognize Black Faces Equally

US government tests find even top-performing facial recognition systems misidentify blacks at rates five to 10 times higher than they do whites.

Source: https://miro.medium.com/max/1400/1*t5-MUraN0EoCntpVTic77Q.png

Bias in AI

Overcoming Racial Bias In AI Systems And Startlingly Even In AI Self-Driving Cars

Racial bias in a medical algorithm favors white patients over sicker black patients

AI expert calls for end to UK use of 'racially biased' algorithms

AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators

Gender bias in AI: building fairer algorithms

Bias in AI: A problem recognized but still unresolved

Amazon, Apple, Google, IBM, and Microsoft worse at transcribing black people's voices than white people's with AI voice recognition, study finds

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

The Week in Tech: Algorithmic Bias Is Bad. Uncovering It Is Good.

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

Artificial Intelligence has a gender bias problem – just ask Siri

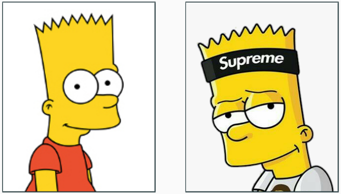
The Best Algorithms Struggle to Recognize Black Faces Equally

US government tests find even top-performing facial recognition systems misidentify blacks at rates five to 10 times higher than they do whites.

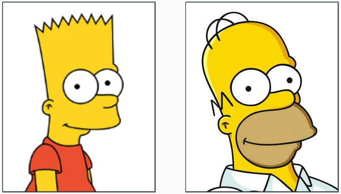
Source: https://miro.medium.com/max/1400/1*t5-MUraN0EoCntpVTic77Q.png

Face Verification Problem

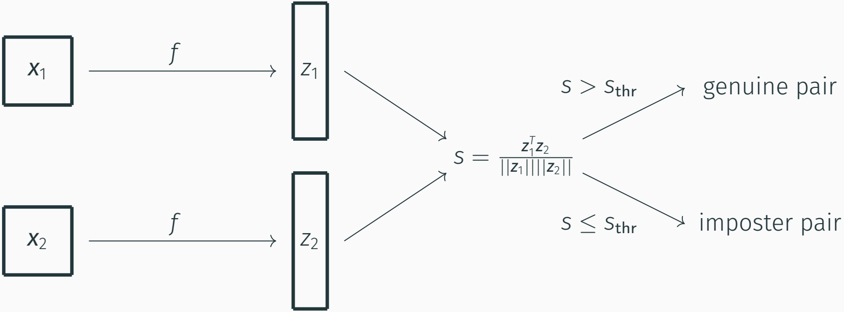
Given two images, decide if it is a genuine/imposter pair.



genuine pair



imposter pair



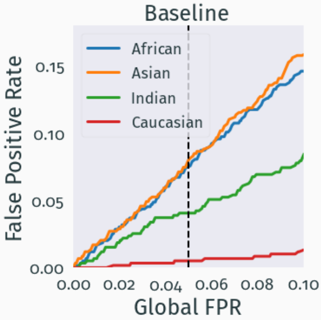
- Measure the similarity between embeddings.
- Threshold to obtain a binary classifier.

Predictive Equality

A binary classifier $\widehat{Y}$ exhibits predictive equality for subgroups g_1 and g_2 if the classifier has equal FPRs for each subgroup,

$$\mathbb{P}_{(x_1,x_2)\sim\mathcal{G}_1}\left(\widehat{Y}=1 \mid Y=0\right)=\mathbb{P}_{(x_1,x_2)\sim\mathcal{G}_2}\left(\widehat{Y}=1 \mid Y=0\right).$$

Results for the FaceNet (Webface) model on the RFW dataset.



Goals and Related Work

Devise a post-hoc method that:

- Improves Accuracy
- Achieves Fairness-calibration
- Achieves Predictive equality (equal FPRs)
- Does not require the sensitive attribute
- Does not require additional training.

Our FairCal method achieves all of the above!

Methods	Improves accuracy	Fairly calibrated	Predictive equality	Does not require sensitive attribute during training	Does not require sensitive attribute at test time	Does not require additional training
AGENDA	✘	✘	✔	✘	✔	✘
PASS	✘	✘	✔	✘	✔	✘
FTC	✘	✘	✔	✘	✔	✘
GST	✔	✘	✔	✘	✘	✔
FSN	✔	✘	✔	✔	✔	✔
FairCal (Ours)	✔	✔	✔	✔	✔	✔

Calibration Stage

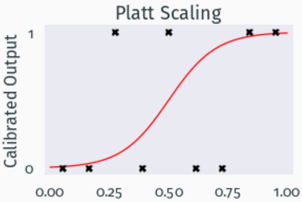
Input: feature embeddings of a set of face images $\mathcal{Z}^{\text{cal}}$

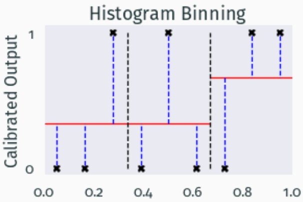
1. Apply the K-means algorithm to $\mathcal{Z}^{\text{cal}}$ partitioning the embedding space $\mathcal{Z}$ into K clusters $\mathcal{Z}_1, \ldots, \mathcal{Z}_k$.
2. Form the K calibration sets

$$S_k^{\text{cal}} = \{s(x_1, x_2) : f(x_1) \in \mathcal{Z}_k \text{ or } f(x_2) \in \mathcal{Z}_k\}, \quad k = 1, \ldots, K$$

3. For $k = 1, \ldots, K$ find a calibration map μ_i such that

$$\mu_k(s(x_1, x_2)) = \mathbb{P}[Y = 1 \mid S = s, f(x_1) \in \mathcal{Z}_k \text{ or } f(x_2) \in \mathcal{Z}_k]$$





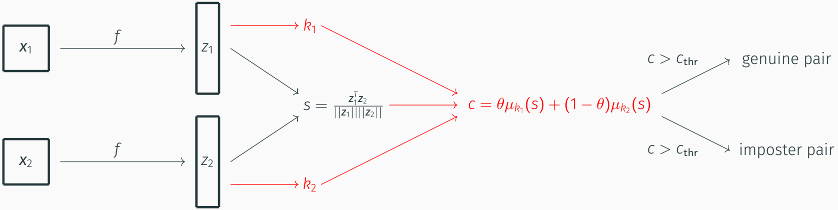
FairCal

Test Stage

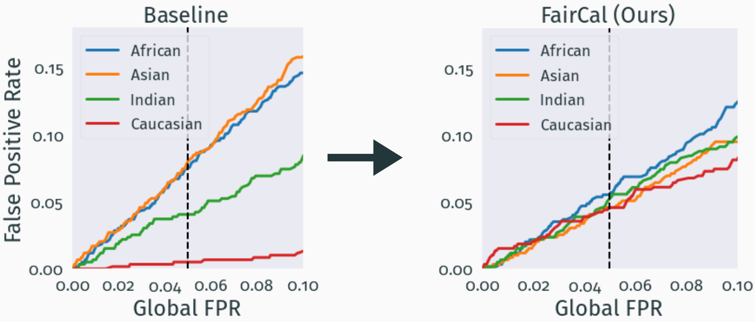
1. Given an image pair (x_1, x_2), compute the cluster of each image feature: k_1 and k_2
2. The model's confidence in it being a genuine pair is

$$c(x_1, x_2) = \theta\mu_{k_1}(s(x_1, x_2)) \; + (1 - \theta)\mu_{k_2}(s(x_1, x_2))$$

where $\theta = \frac{|S^{\text{cal}}_{k1}|}{|S^{\text{cal}}_{k1}| + |S^{\text{cal}}_{k2}|}$ is the relative population fraction of the two clusters.



Results - Predictive Equality



Results - Predictive Equality

Comparison of subgroup FPRs in terms of AAD, MAD, STD.

		RFW						BFW					
		FaceNet (VGGFace2)			FaceNet (Webface)			FaceNet (Webface)			ArcFace		
	(↓)	AAD	MAD	STD	AAD	MAD	STD	AAD	MAD	STD	AAD	MAD	STD
0.1% FPR	Baseline	0.10	0.15	0.10	0.14	0.26	0.16	0.29	1.00	0.40	0.12	0.30	0.15
	AGENDA	0.11	0.20	0.13	0.12	0.23	0.14	0.14	0.40	0.18	**0.09**	0.23	**0.11**
	PASS	0.11	0.18	0.12	0.11	0.18	0.12	0.36	1.21	0.49	0.12	0.29	0.14
	FTC	0.10	0.15	0.11	0.12	0.23	0.14	0.24	0.74	0.32	**0.09**	**0.20**	**0.11**
	GST	0.13	0.24	0.15	**0.09**	**0.16**	**0.10**	0.13	0.35	0.16	0.10	0.24	0.12
	FSN	0.10	0.18	0.11	0.11	0.23	0.13	**0.09**	0.20	**0.11**	0.11	0.28	0.14
	FairCal (Ours)	**0.09**	**0.14**	**0.10**	**0.09**	**0.16**	**0.10**	**0.09**	**0.20**	0.11	0.11	0.31	0.15
1% FPR	Baseline	0.68	1.02	0.74	0.67	1.23	0.79	2.42	7.48	3.22	0.72	1.51	0.85
	AGENDA	0.73	1.14	0.81	0.73	1.08	0.78	1.21	3.09	1.51	0.65	1.78	0.84
	PASS	0.89	1.52	1.01	0.68	0.99	0.73	3.30	10.18	4.34	0.72	2.00	0.93
	FTC	0.60	0.91	0.66	0.54	1.05	0.66	1.94	5.74	2.57	0.54	**1.04**	0.61
	GST	0.52	0.92	0.60	0.30	0.57	0.37	1.05	3.01	1.38	**0.44**	1.13	**0.56**
	FSN	0.37	0.68	0.46	0.35	0.61	0.40	0.87	2.19	1.05	0.55	1.27	0.68
	FairCal (Ours)	**0.28**	**0.46**	**0.32**	**0.29**	**0.57**	**0.35**	**0.80**	**1.79**	**0.95**	0.63	1.46	0.78

AAD - Average Absolute Deviation; MAD - Maximum Absolute Deviation (MAD); STD - Standard Deviation

Results

Accuracy

(↑)	RFW						BFW					
	FaceNet (VGGFace2)			FaceNet (Webface)			FaceNet (Webface)			ArcFace		
	AUROC	TPR @ 0.1% FPR	TPR @ 1% FPR	AUROC	TPR @ 0.1% FPR	TPR @ 1% FPR	AUROC	TPR @ 0.1% FPR	TPR @ 1% FPR	AUROC	TPR @ 0.1% FPR	TPR @ 1% FPR
Baseline	88.26	18.42	34.88	83.95	11.18	26.04	96.06	33.61	58.87	97.41	86.27	90.11
AGENDA	76.83	8.32	18.01	74.51	6.38	14.98	82.42	15.95	32.51	95.09	69.61	79.67
PASS	86.96	13.67	29.30	81.44	7.34	20.93	92.27	17.21	38.32	96.55	77.38	85.26
FTC	86.46	6.86	23.66	81.61	4.65	18.40	93.30	13.60	43.09	96.41	82.09	88.24
GST	89.57	22.61	40.72	84.88	17.34	31.56	96.59	44.49	66.71	96.89	86.13	89.70
FSN	90.05	23.01	40.21	85.84	17.33	32.80	96.77	**47.11**	68.92	97.35	86.19	90.06
FairCal (Ours)	**90.58**	**23.55**	**41.88**	**86.71**	**20.64**	**33.13**	**96.90**	46.74	**69.21**	**97.44**	**86.28**	**90.14**

Fairness-Calibration

(↓)	RFW								BFW							
	FaceNet (VGGFace2)				FaceNet (Webface)				FaceNet (Webface)				ArcFace			
	Mean	AAD	MAD	STD	Mean	AAD	MAD	STD	Mean	AAD	MAD	STD	Mean	AAD	MAD	STD
Baseline	6.37	2.89	5.73	3.77	5.55	2.48	4.97	2.91	6.77	3.63	5.96	4.03	2.57	1.39	2.94	1.63
AGENDA	7.71	3.11	6.09	3.86	5.71	2.37	4.28	2.85	13.21	6.37	12.91	7.55	5.14	2.48	5.92	3.04
PASS	8.09	2.40	4.10	2.83	7.65	3.36	5.34	3.85	13.16	5.25	9.58	6.12	3.69	2.01	4.24	2.37
FTC	5.69	2.32	4.51	2.95	4.73	1.93	3.86	2.28	6.64	2.80	5.61	3.27	2.95	1.48	3.03	1.74
GST	2.34	0.82	1.58	0.98	3.24	1.21	1.93	1.34	3.09	1.45	2.80	1.65	3.34	1.81	4.21	2.19
FSN	1.43	0.35	0.57	0.40	2.49	0.84	1.19	0.91	**2.76**	1.38	2.67	1.60	2.65	1.45	3.23	1.71
FairCal (Ours)	**1.37**	**0.28**	**0.50**	**0.34**	**1.75**	**0.41**	**0.64**	**0.45**	3.09	**1.34**	**2.48**	**1.55**	**2.49**	**1.30**	**2.68**	**1.52**

AAD - Average Absolute Deviation; MAD - Maximum Absolute Deviation (MAD); STD - Standard Deviation

Thank You!

Contact: tiago.saldanhasalvador@mcgill.ca