Characterizing Bias in Classifiers using Generative Models

Daniel McDuff   Shuang Ma   Yale Song   Ashish Kapoor
damcduff@microsoft.com   shuangma@buffalo.edu   yalesong@microsoft.com   akapoor@Microsoft.com

Data and Code is available at:
https://github.com/danmcduff/characterizingBias

Problem & Motivation

- Models are often biased because the data used to train them is biased.
- Existing approaches rely on oracles labeling which are ultimately limited.
- We propose a simulation-based approach for interrogating classifiers in a systematic manner.

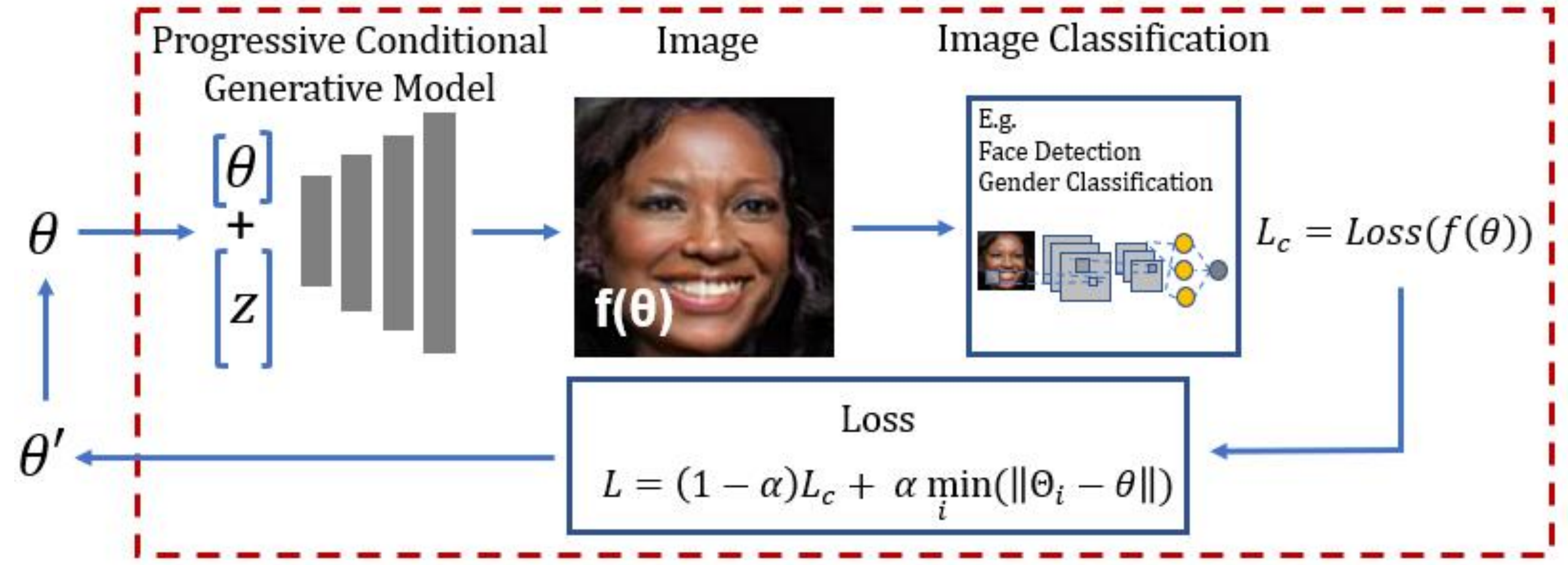


Figure 1. The system pipeline.

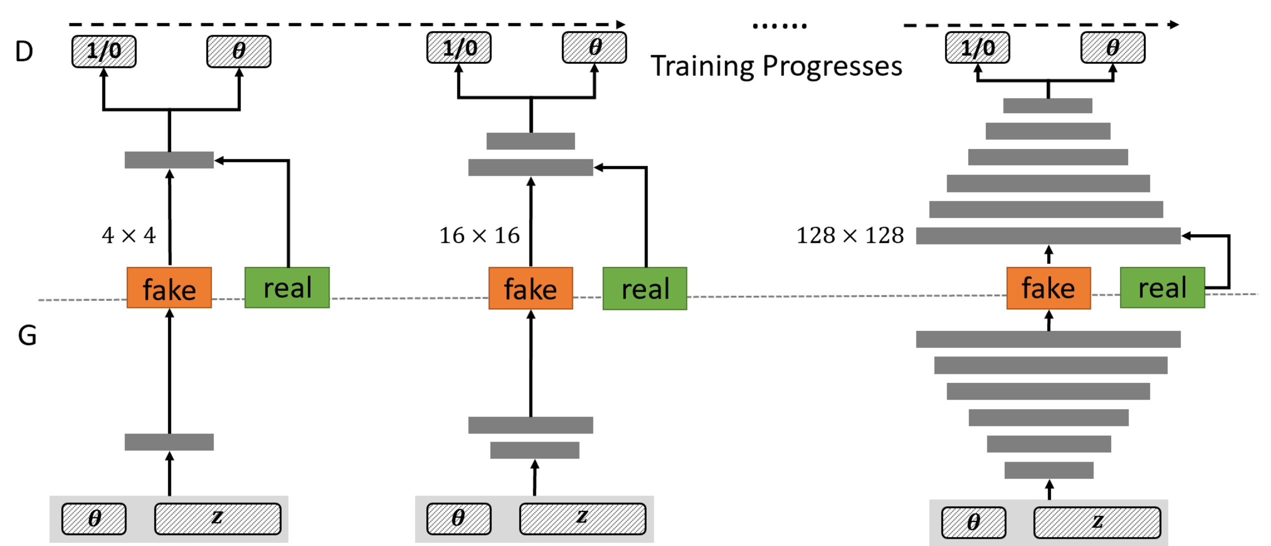
Contribution

1. Presenting an approach for conditionally generating synthetic face images.
2. Showing how synthetic data can be used to efficiently identify bias.
3. Proposing a Bayesian Optimization sampling procedure to identify bias more efficiently.
4. Releasing the dataset, model and code.

Approach

We first use a *Generative Model* to synthesize face images and then apply *Bayesian Optimization* to find errors in the target classifier.

Image Generation



$$\mathcal{L}_G = -\mathbb{E}_{z,\theta}[\log D(G(z,\theta))]$$

$$\mathcal{L}_D = -\mathbb{E}[\log D(x)] - \mathbb{E}_{z,\theta}[\log D(G(z,\theta))] - \mathbb{E}_{z,\theta}[\log C(G(z,\theta))]$$

$$\mathcal{L}_{adv} = \min_G \max_D \mathcal{L}_G + \mathcal{L}_D$$

Bayesian Optimization

Classification Loss: $L_c = Loss(f(\theta))$

L_c reflects the misclassification cost when applying θ. $\theta = [race; gender]$

Composite Loss:

$$L = (1-\alpha)L_c + \alpha \min_i \|\Theta_i - \theta\|$$

The second term encourages exploration and prioritizes sampling a diverse set of images.

Data

Region	Country	People M	People W	Frames M	Frames W
Black	Nigerian	81	28	768	467
	Kenya	11	5	91	49
	S. Africa	136	102	1641	1984
	Total	228	135	2500	2500
S Asian	India	142	83	2108	2267
	Sri Lanka	1	2	11	7
	Pakistan	19	11	381	226
	Total	162	96	2500	2500
White	Australia	175	121	2500	2500
	Total	175	121	2500	2500
NE Asian	Japan	105	89	930	1421
	China	105	46	789	447
	S. Korea	29	12	464	251
	Hong Kong	36	28	317	381
	Total	275	175	2500	2500

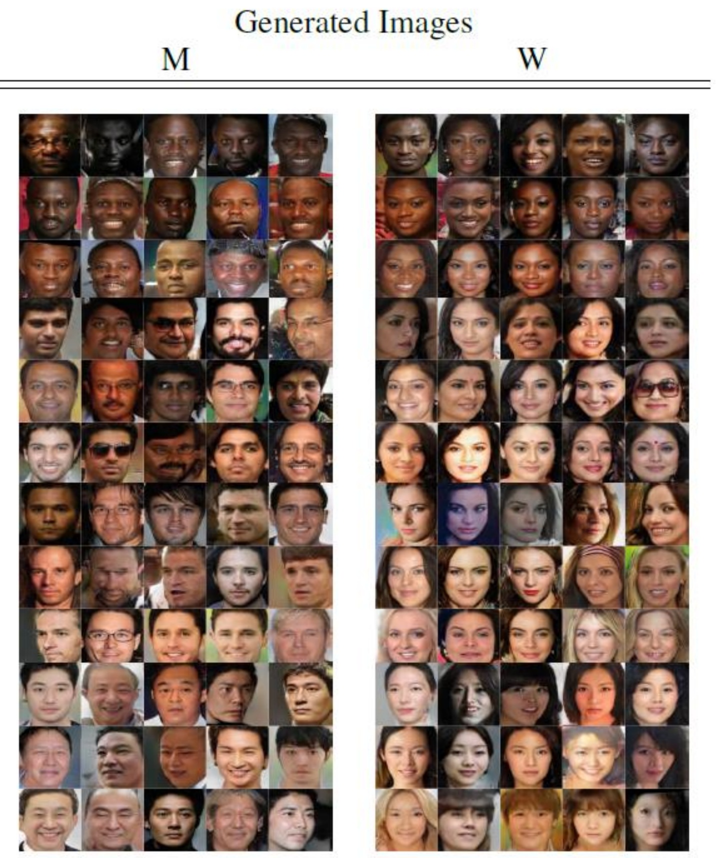


Table 1. The number of people and images we sampled from (by country and gender) to train our generation model.

We sampled a balanced subset from MS-CELEB-1M.

1. Utilize Google Search API and NLTK to extract nationality and gender information.
2. Sample evenly distributed images only from countries with more homogeneous demographics.

Evaluation

✓ Validation of Image Generation

FID Score

	Black		White		North East Asian		South Asian	
	Male	Female	Male	Female	Male	Female	Male	Female
Ours	8.10	8.14	8.08	7.70	8.01	8.00	8.06	8.10
StyleGAN	7.70	7.92	7.68	7.80	7.76	7.80	7.94	7.66

✓ Classifier Interrogation

API	Task	All	Black	S. Asian	NE Asian	White	Men	Women
IBM	Face Det.	8.05	**16.9**	**7.63**	3.96	3.80	**11.3**	2.27
	Gender Class.	8.26	**9.00**	2.13	**20.0**	1.87	**15.8**	0.27
SE	Face Det.	0.13	0.00	0.00	0.53	0.00	0.21	0.00
	Gender Class.	2.84	3.39	0.74	**5.85**	1.38	5.14	0.00

Table 2: Face and gender classification error rates.

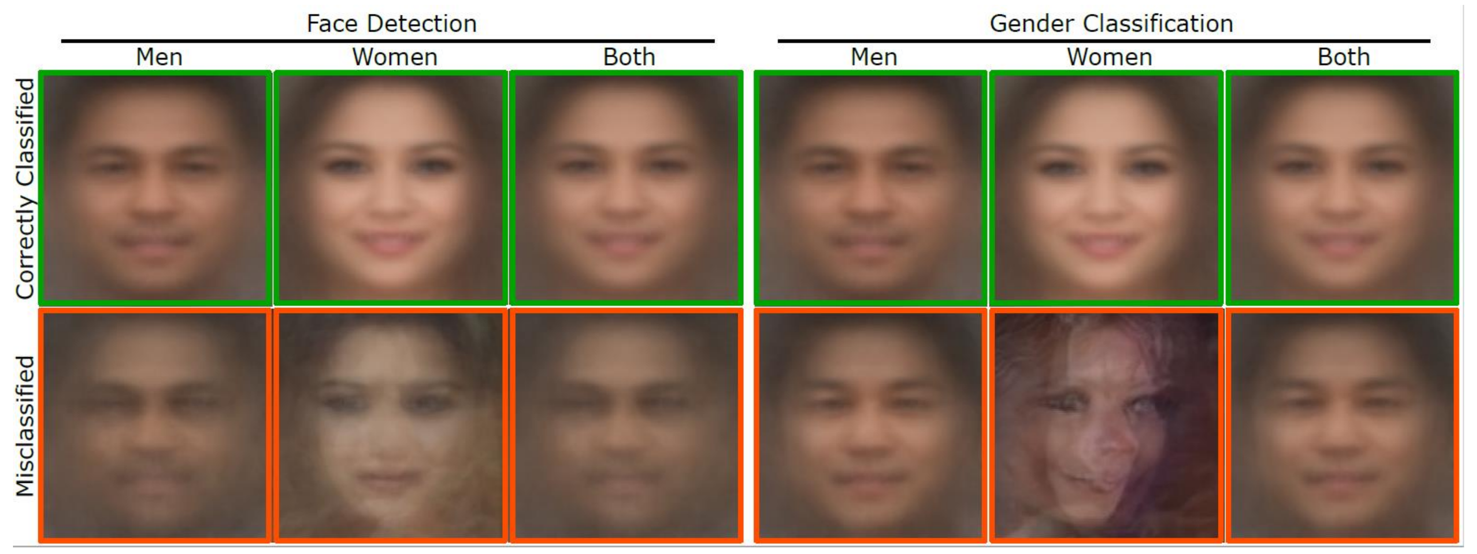


Figure 2. Mean faces for correct classifications and incorrect classifications.

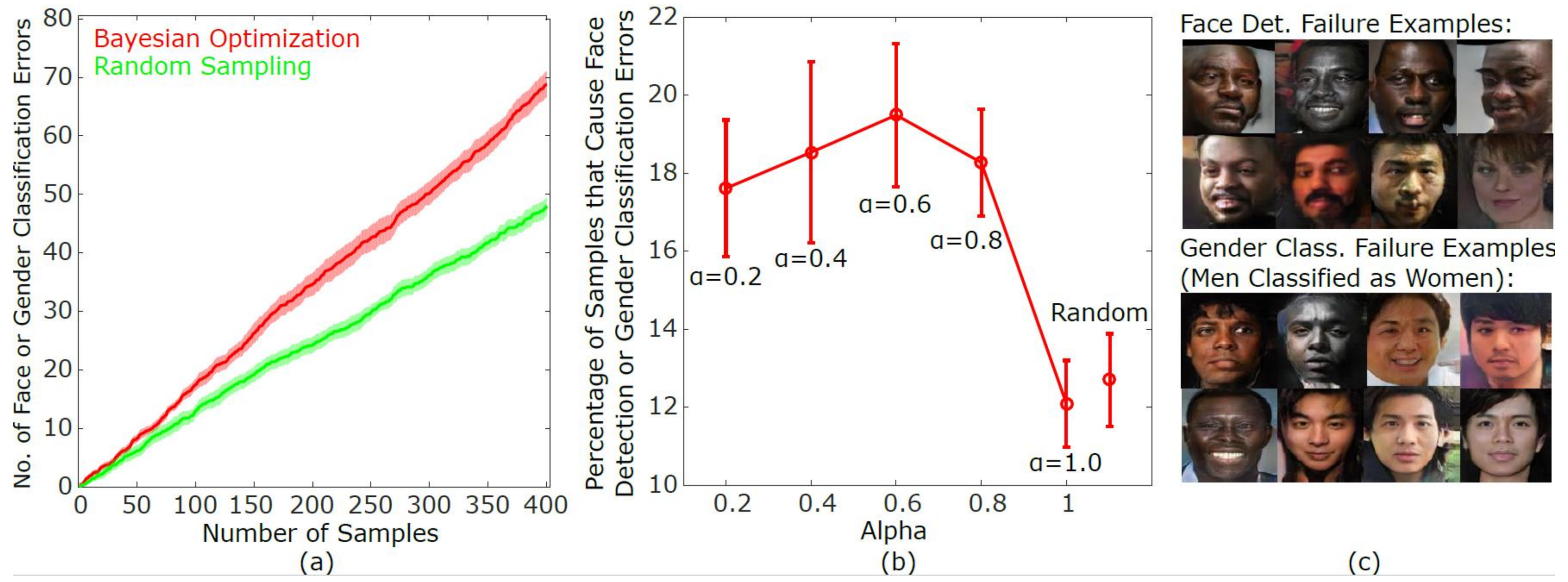


Figure 3: a) Sample efficiency of finding samples that were misclassified using random sampling and Bayesian Optimization with =1. b) Percentage of images that cause classifier failures (y-axis) as we vary the value of α. c) Qualitative examples of failure cases.