Fair Generative Modeling via Weak Supervision

Kristy Choi   Aditya Grover   Trisha Singh Rui Shu Stefano Ermon

ICML 2020

Motivation

Figure 1. Samples from a baseline BigGAN that reflect the gender bias underlying the true data distribution in CelebA. All faces above the orange line (67%) are classified as female, while the rest are labeled as male (33%).

biased dataset will result that generative model have biased performance

Setup

two dataset :

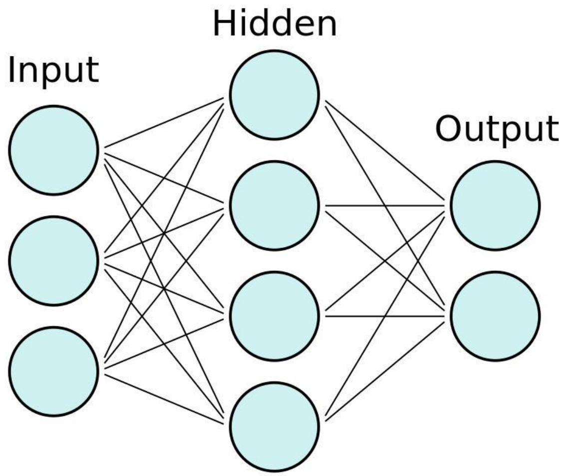
$\mathcal{D}_{\text{ref}} \longrightarrow$ A small unbiased dataset

$\mathcal{D}_{\text{bias}} \longrightarrow$ A large biased dataset

Goal: generated data Pdata $p_{\text{data}} = p_{\text{ref}}$

Method: Importance Reweighting

binary classification



$$Y = 1 \longrightarrow \mathcal{D}_{\text{ref}}$$

$$Y = 0 \longrightarrow \mathcal{D}_{\text{bias}}$$

Loss:

$$NCE(c) := \frac{1}{\gamma+1}\mathbb{E}_{p_{\text{ref}}(\mathbf{x})}[\log c(Y=1|\mathbf{x})] + \frac{\gamma}{\gamma+1}\mathbb{E}_{p_{\text{bias}}(\mathbf{x})}[\log c(Y=0|\mathbf{x})].$$

Reweighting:

$$w(\mathbf{x}) = \frac{p_{\text{ref}}(\mathbf{x})}{p_{\text{bias}}(\mathbf{x})} = \gamma \frac{c^*(Y=1|x)}{1-c^*(Y=1|x)} \qquad (5)$$

Algorithm 1 Learning Fair Generative Models

Input: $\mathcal{D}_{\text{bias}}, \mathcal{D}_{\text{ref}}$, Classifier and Generative Model Architectures & Hyperparameters

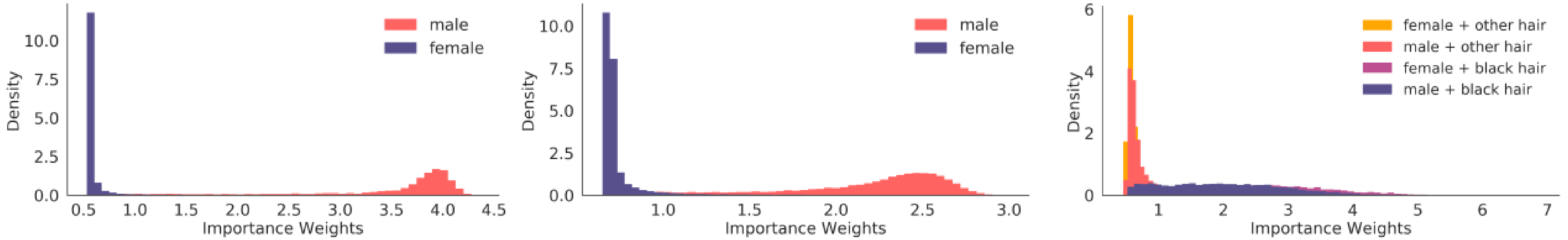
Output: Generative Model Parameters θ

```
▷ Phase 1: Estimate importance weights
Learn binary classifier c for distinguishing (D_bias, Y = 0) vs. (D_ref, Y = 1)
Estimate importance weight ŵ(x) ← c(Y=1|x) / c(Y=0|x) for all x ∈ D_bias (using Eq. 5)
Set importance weight ŵ(x) ← 1 for all x ∈ D_ref

▷ Phase 2: Minibatch gradient descent on θ based on weighted loss
Initialize model parameters θ at random
Set full dataset D ← D_bias ∪ D_ref
while training do
    Sample a batch of points B from D at random
    Set loss L(θ; D) ← 1/|B| Σ_{x_i ∈ B} ŵ(x_i) ℓ(x_i, θ)
    Estimate gradients ∇_θ L(θ; D) and update parameters θ based on optimizer update rule
end while
return θ
```

$$w(\mathbf{x}) = \frac{p_{\text{ref}}(\mathbf{x})}{p_{\text{bias}}(\mathbf{x})} = \gamma \frac{c^*(Y=1|x)}{1 - c^*(Y=1|x)} \tag{5}$$

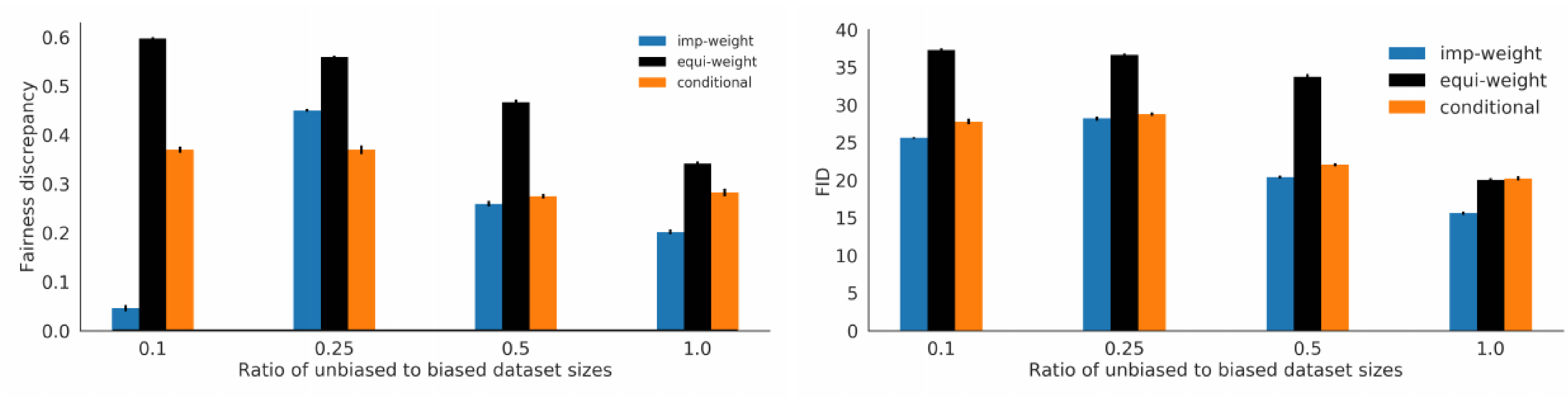
Experiments



(a) single, `bias`=0.9 (b) single, `bias`=0.8 (c) multi

Figure 2. Distribution of importance weights for different latent subgroups. On average, The underrepresented subgroups are upweighted while the overrepresented subgroups are downweighted.

(a) Samples generated via importance reweighting with subgroups separated by the orange line. For the 100 samples above, the classifier concludes 52 females and 48 males.

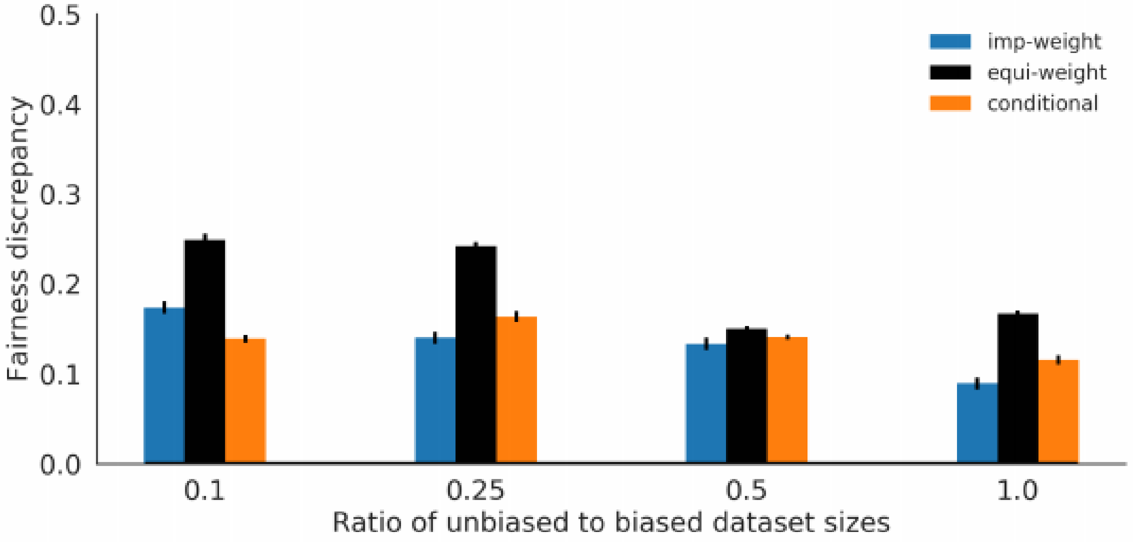


(b) Fairness Discrepancy

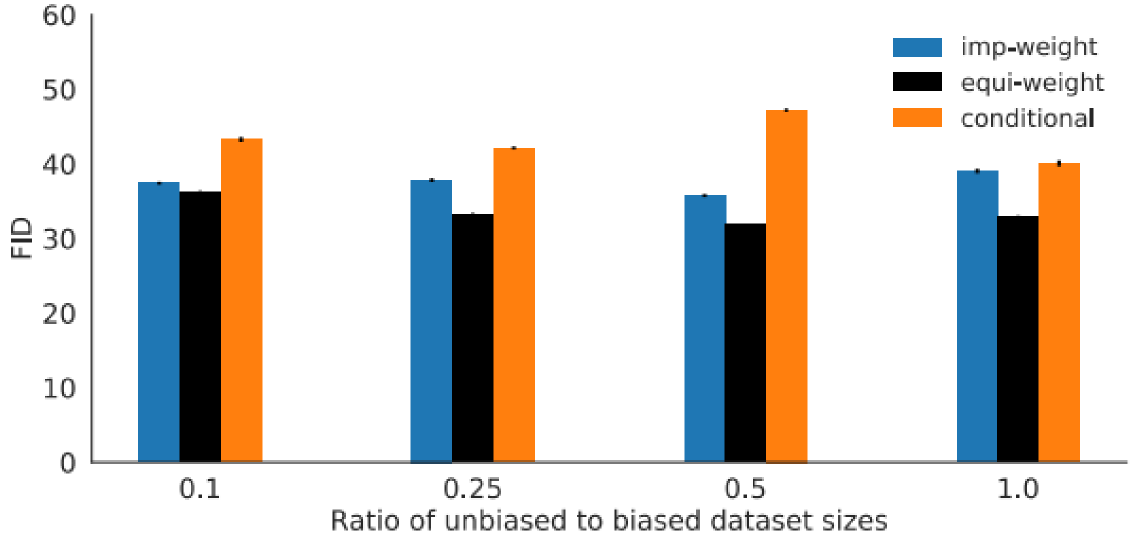
(c) FID

Figure 3. Single-Attribute Dataset Bias Mitigation for `bias`=0.9. Lower discrepancy and FID is better. Standard error in (b) and (c) over 10 independent evaluation sets of 10,000 samples each drawn from the models. We find that on average, `imp-weight` outperforms the `equi-weight` baseline by 49.3% and the `conditional` baseline by 25.0% across all reference dataset sizes for bias mitigation.

(a) Samples generated via importance reweighting. For the 100 samples above, the classifier concludes 37 females and 20 males without black hair, 22 females and 21 males with black hair.



(b) Fairness Discrepancy



(c) FID

Figure 4. Mult-Attribute Dataset Bias Mitigation. Standard error in (b) and (c) over 10 independent evaluation sets of 10,000 samples each drawn from the models. Lower discrepancy and FID is better. We find that on average, `imp-weight` outperforms the `equi-weight` baseline by 32.5% and the `conditional` baseline by 4.4% across all reference dataset sizes for bias mitigation.