Characterizing Bias in Classifiers using Generative Models

McDuff et al., NeurIPS 2019

Presenter: Sai Anuroop Kesanapalli

CSCI 535: Multimodal Probabilistic Learning of Human Communication
Spring 2024

University of Southern California

Table of Contents

Motivation

- Models learnt from real-world data are often biased because the data on which they are trained is biased.

Motivation

- Models learnt from real-world data are often biased because the data on which they are trained is biased.
- To characterize these biases, existing approaches rely on human annotators labeling real-life examples to identify the “blind spots” of the classifiers.

Motivation

- Models learnt from real-world data are often biased because the data on which they are trained is biased.
- To characterize these biases, existing approaches rely on human annotators labeling real-life examples to identify the "blind spots" of the classifiers.
  - Labor intensive / costly
  - Exisiting real-world images are finite

Motivation

- Models learnt from real-world data are often biased because the data on which they are trained is biased.
- To characterize these biases, existing approaches rely on human annotators labeling real-life examples to identify the "blind spots" of the classifiers.
  - Labor intensive / costly
  - Exisiting real-world images are finite
- **Core idea:** Why not use a simulation-based approach using generative adversarial models in a systematic manner? [5]

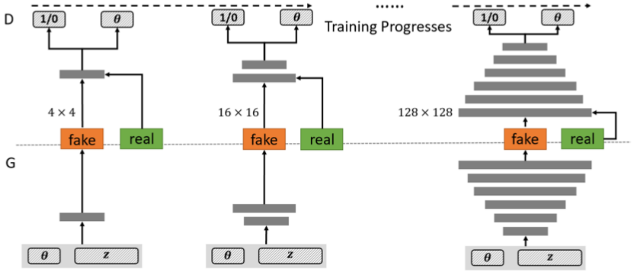
Table of Contents

Approach

- Devise a systematically controllable image generation model.

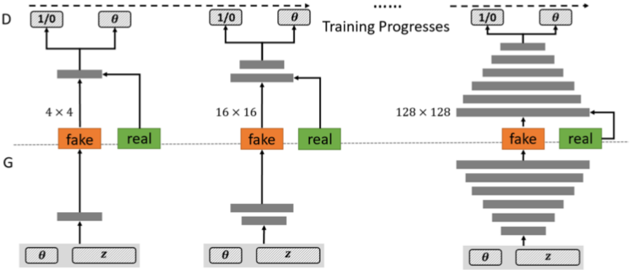
Approach

- Devise a systematically controllable image generation model.
  - **Model:** Progressively growing conditional GAN [4].



Approach

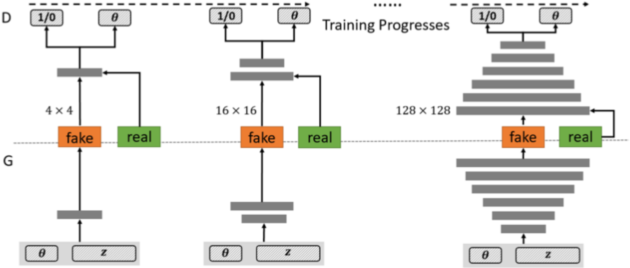
- Devise a systematically controllable image generation model.
  - **Model:** Progressively growing conditional GAN [4].



  - **Dataset:** $\{x, \theta\}$ is a curated dataset, where x is a face image, and $\theta := [r; g]$ is a one-hot representation that specifies both the race r and gender g of the subject in the image.
  - **Input:** $p_z(z)$ is a prior noise input where $z \sim \mathcal{U}(0, 1)$. Concatenated z and θ is the input to model.

Approach

- Devise a systematically controllable image generation model.
  - **Model:** Progressively growing conditional GAN [4].



  - **Dataset:** $\{x, \theta\}$ is a curated dataset, where x is a face image, and $\theta := [r; g]$ is a one-hot representation that specifies both the race r and gender g of the subject in the image.
  - **Input:** $p_z(z)$ is a prior noise input where $z \sim \mathcal{U}(0, 1)$. Concatenated z and θ is the input to model.
  - **Loss:**

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

Approach

- Combine this generation model with a Bayesian Optimization process to find out sets of diverse examples that would lead to misclassification.

Approach

- Combine this generation model with a Bayesian Optimization process to find out sets of diverse examples that would lead to misclassification.
  - **Objective:**

$$\boxed{\text{maximize } L = (1-\alpha)L_c + \alpha \min_i ||\Theta_i - \theta||}$$

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

Approach

- Combine this generation model with a Bayesian Optimization process to find out sets of diverse examples that would lead to misclassification.
  - **Objective:**

$$\boxed{\text{maximize } L = (1-\alpha)L_c + \alpha \min_i ||\Theta_i - \theta||}$$

    The second term encourages *exploration* and prioritizes sampling a diverse set of images.
  - **Process:** L is modeled as a Gaussian process. Modeling as a GP helps quantify uncertainty around the predictions, which in turn is used to efficiently explore the parameter space θ in order to identify the spots that satisfy the search criterion.

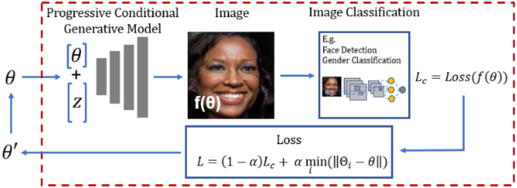


Table of Contents

Experiments

Setup

- **Data:** MS-CELEB-1M [2], a large image dataset with a training set containing 100K different people and approximately 10 million images[1].

[1]Some demographic caveats included, details in the paper

- **Data:** MS-CELEB-1M [2], a large image dataset with a training set containing 100K different people and approximately 10 million images[1].
- **Validation of Image Generation:** Generated a uniform sample of 50 images, at 128×128 resolution, from each race and gender (total $50 \times 4 \times 2 = 400$ images) and recruited five participants on MTurk to label the gender of the face in each image and the quality of the image. Additionally, FID scores computed for each region and gender for the conditional PG-GAN and compared with StyleGAN [3].

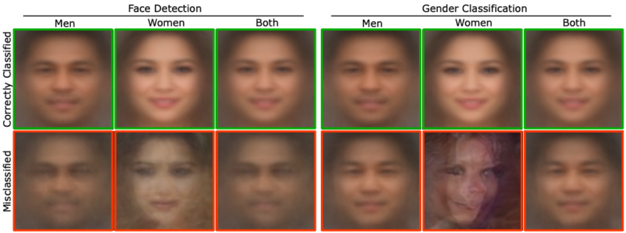
[1] Some demographic caveats included, details in the paper

- **Data:** MS-CELEB-1M [2], a large image dataset with a training set containing 100K different people and approximately 10 million images[1].
- **Validation of Image Generation:** Generated a uniform sample of 50 images, at 128×128 resolution, from each race and gender (total $50 \times 4 \times 2 = 400$ images) and recruited five participants on MTurk to label the gender of the face in each image and the quality of the image. Additionally, FID scores computed for each region and gender for the conditional PG-GAN and compared with StyleGAN [3].
- **Classifier Interrogation:** IBM and SightEngine commercial classifiers used. 400 images at 128×128 resolution sampled and used for interrogation.

[1] Some demographic caveats included, details in the paper

Experiments

Results

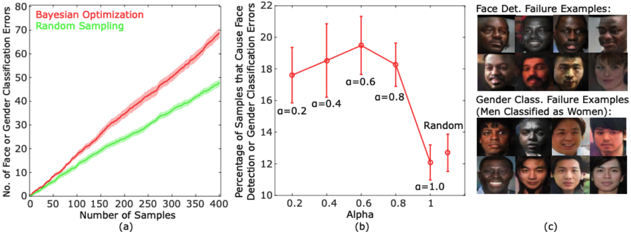


Experiments
Results

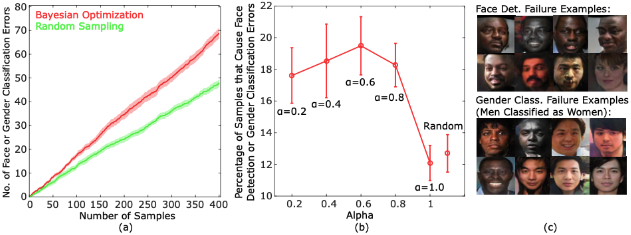
- Missed faces had darker skin tones and gender classification was considerably less accurate on people from NE Asia.
- Men were more frequently misclassified as women.

Experiments

Results



[2] Results based on IBM API



- (a) Bayesian Optimization is better than random sampling in finding these "blind spots".
- (b) $\alpha = 0.6$ gives the sets of samples that cause most misclassifications[2].
- (c) Examples of images that resulted in errors.

[2] Results based on IBM API

Table of Contents

Discussion

Limitations –

- Although very detailed, this work is limited by the capability of conditional PG-GAN.
- Bayesian Optimization is only "linearly" better than random sampling.
- Other types of loss, in addition to diversity loss, can be included in the cumulative loss for Bayesian Optimization that may lead to better results.

Discussion

Limitations –

- Although very detailed, this work is limited by the capability of conditional PG-GAN.
- Bayesian Optimization is only "linearly" better than random sampling.
- Other types of loss, in addition to diversity loss, can be included in the cumulative loss for Bayesian Optimization that may lead to better results.

Food for thought –

- This work addresses sample selection bias. For example, *Gender Shades* project [1] gives an alarming overview of how bad these biases are!
- Other forms of biases exist that may manifest as real-world biases.
  - *Implicit biases* in optimizers such as SGD: directional [7], minimum-norm [6].
  - *Hallucinations* of generative models: inaccurate and misleading outputs.

Table of Contents

References I

[1] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Conference on fairness, accountability and transparency*. PMLR. 2018, pp. 77–91.

[2] Yandong Guo et al. "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer. 2016, pp. 87–102.

[3] Tero Karras, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 4401–4410.

References II

[4] Tero Karras et al. "Progressive Growing of GANs for Improved Quality, Stability, and Variation". In: *International Conference on Learning Representations*. 2018.

[5] Daniel McDuff et al. "Characterizing bias in classifiers using generative models". In: *Advances in neural information processing systems* 32 (2019).

[6] Jiyoung Park, Ian Pelakh, and Stephan Wojtowytsch. "Minimum norm interpolation by perceptra: Explicit regularization and implicit bias". In: *Advances in Neural Information Processing Systems* 36 (2023).

[7] Jingfeng Wu et al. "Direction Matters: On the Implicit Bias of Stochastic Gradient Descent with Moderate Learning Rate". In: *International Conference on Learning Representations*. 2020.

Thank you!