DRO vs Fairness

Roman Silen, Jared Gridley, Dan Stevens, Cole Mediratta, William Hawkins

Overview

- Introduction
  - Motivation and Definitions
- Paper Overviews
  - Fairness - Fair Attribute Classification through Latent Space De-biasing (Ramaswamy et al., 2021)
  - DRO - Distributionally Robust Neural Networks For Group Shifts (Sagawa et al., 2020)
- Fairness Experiment and Results
- DRO Experiment and Results
- Fairness + DRO Experiment and Results
- Conclusions

Overview

- Introduction
  - Motivation and Definitions
- Paper Overviews
  - Fairness - Fair Attribute Classification through Latent Space De-biasing (Ramaswamy et al., 2021)
  - DRO - Distributionally Robust Neural Networks For Group Shifts (Sagawa et al., 2020)
- Fairness Experiment and Results
- DRO Experiment and Results
- Fairness + DRO Experiment and Results
- Conclusions

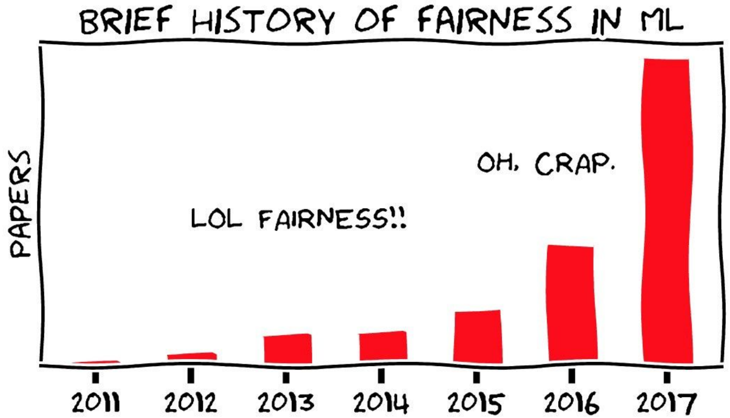
Motivation

- Robustness and Fairness are two incredibly important metrics in machine learning.

- Keeping models fair and robust ensures that they are able to effectively generalize and be impartial to create overall trustworthiness in our machine learning algorithms.

- Our novel idea a with solving **<u>both</u>** problems with one algorithm. We would like to create a fair and generalizable algorithm.

Definitions



Distributionally Robust Optimization (DRO) - The collection of optimization methods to maximize robustness against learning “spurious correlations” among features in data.

Fairness - A concept in machine learning that aims to create a model that performs as optimally as possible whilst simultaneously treating data points fairly based on certain attributes.

Data: CelebA[3]

- 10,177 number of identities
- 202,599 number of face images
- 5 landmark locations, 40 binary attributes annotations per image
  - Ex. {blond, not blond} and {male, not male}

[3] Liu et al. (2015). Deep Learning Face Attributes in the Wild.

Overview

- Introduction
  - Motivation and Definitions
- Paper Overviews
  - Fairness - Fair Attribute Classification through Latent Space De-biasing (Ramaswamy et al., 2021)
  - DRO - Distributionally Robust Neural Networks For Group Shifts (Sagawa et al., 2020)
- Fairness Experiment and Results
- DRO Experiment and Results
- Fairness + DRO Experiment and Results
- Conclusions

Paper Overview - Fairness

- Ramaswamy et al. (2021). Fair Attribute Classification through Latent Space De-biasing[1]
  - Conference on Computer Vision and Pattern Recognition (CVPR) 2021

PROBLEM: Bad data sets (or even good ones) can lead to models learning spurious relationships between attributes.

Possible Solution: Use a GAN to generate new images for the data set to offset potential biases, but even this inherits biases from original dataset

Novel Idea: Use latent vector perturbation method to debias the generated images, producing a more fair dataset.

arXiv:2012.01469v3 [cs.CV] 2 Apr 2021

Fair Attribute Classification through Latent Space De-biasing

Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky
Princeton University
{vr23, suhk, olgarus}@cs.princeton.edu

Abstract

Fairness in visual recognition is becoming a prominent and critical topic of discussion as recognition systems are deployed at scale in the real world. Models trained from data in which target labels are correlated with protected attributes (e.g., gender, race) are known to learn and exploit those correlations. In this work, we introduce a method for training accurate target classifiers while mitigating biases that stem from these correlations. We use GANs to generate realistic-looking images, and perturb these images in the underlying latent space to generate training data that is balanced for each protected attribute. We augment the original dataset with this generated data, and empirically demonstrate that target classifiers trained on the augmented dataset exhibit a number of both quantitative and qualitative benefits. We conduct a thorough evaluation across multiple target labels and protected attributes in the CelebA dataset, and provide an in-depth analysis and comparison to existing literature in the space. Code can be found at https://github.com/princetonvisualai/gan-debiasing.

1. Introduction

Large-scale supervised learning has been the driving force behind advances in visual recognition. Recently, however, there has been a growing number of concerns about the disparate impact of these visual recognition systems. Face recognition systems trained from datasets with an under-representation of certain racial groups have exhibited lower accuracy for those groups [9]. Activity recognition models trained on datasets with high correlations between the activity and the gender expression of the depicted person have over-amplified those correlations [46]. Computer vision systems are statistical models that are trained to maximize accuracy on the majority of examples, and they do so by exploiting the most discriminative cues in a dataset, potentially learning spurious correlations. In this work, we introduce a new framework for training computer vision models that aims to mitigate such concerns, illustrated in Figure 1.

One proposed path for building 'fairer' computer vision

Real world data | Paired augmentation
hat / no hat
hat is correlated with glasses
hat is no longer correlated with glasses

Figure 1: Training a visual classifier for an attribute (e.g., hat) can be complicated by correlations in the training data. For example, the presence of hats can be correlated with the presence of glasses. We propose a dataset augmentation strategy using Generative Adversarial Networks (GANs) that successfully removes this correlation by adding or removing glasses from existing images, creating a balanced dataset.

systems is through a 'fairer' data collection process. Works such as [9, 43] propose techniques for better sampling data to more accurately represent all people. Creating a perfectly balanced dataset, however, is infeasible in many cases. With the advances in Generative Adversarial Networks (GANs) [17], several works propose using generated data to augment real-world datasets [12, 35, 42]. These methods have been growing in computational and algorithmic complexity (e.g., [35, 42] adding multiple loss functions to GAN training), necessitating access to a sufficient number of inter-sectional real-world samples. In contrast, we demonstrate a simple and novel data augmentation technique that uses a single GAN trained on a biased real-world dataset.

Illustrative example: Consider our example from Figure 1. Our goal is to train a visual recognition model that recognizes the presence of an attribute, such as wearing a hat. Suppose in the real world wearing a hat is correlated with wearing glasses—for example, because people often wear both hats and sunglasses outside and take them off inside. This correlation may be reflected in the training data, and a classifier trained to recognize a hat may rely on the presence of glasses. Consequently, the classifier may fail to recognize a hat in the absence of glasses, and vice versa.

We propose using a GAN to generate more images with hats but not glasses and images with glasses but not hats, such that WearingHat is de-correlated from Glasses in the training data, by making perturbations in the latent space. Building on work by Denton et al. [14], which demonstrates

1

Paper Overview - DRO

- Sagawa et al. (2020). Distributionally Robust Neural Networks For Group Shifts[2]

  International Conference on Learning Representations (ICLR) 2020

Key Takeaway:

Using the DRO algorithm with a strong L2 penalty we can substantially increase the worst group accuracy (measure of distributional robustness) by about 10-40 percent.

Published as a conference paper at ICLR 2020

DISTRIBUTIONALLY ROBUST NEURAL NETWORKS FOR GROUP SHIFTS: ON THE IMPORTANCE OF REGULARIZATION FOR WORST-CASE GENERALIZATION

Shiori Sagawa*
Stanford University
ssagawa@cs.stanford.edu

Pang Wei Koh*
Stanford University
pangwei@cs.stanford.edu

Tatsunori B. Hashimoto
Microsoft
tahashim@microsoft.com

Percy Liang
Stanford University
pliang@cs.stanford.edu

ABSTRACT

Overparameterized neural networks can be highly accurate *on average* on an i.i.d. test set yet consistently fail on atypical groups of the data (e.g., by learning spurious correlations that hold on average but not in such groups). Distributionally robust optimization (DRO) allows us to learn models that instead minimize the *worst-case* training loss over a set of pre-defined groups. However, we find that naively applying group DRO to overparameterized neural networks fails: these models can perfectly fit the training data, and any model with vanishing average training loss also already has vanishing worst-case training loss. Instead, the poor worst-case performance arises from poor *generalization* on some groups. By coupling group DRO models with increased regularization—a stronger-than-typical ℓ_2 penalty or early stopping—we achieve substantially higher worst-group accuracies, with 10–40 percentage point improvements on a natural language inference task and two image tasks, while maintaining high average accuracies. Our results suggest that regularization is important for worst-group generalization in the overparameterized regime, even if it is not needed for average generalization. Finally, we introduce a stochastic optimization algorithm, with convergence guarantees, to efficiently train group DRO models.

1 INTRODUCTION

Machine learning models are typically trained to minimize the average loss on a training set, with the goal of achieving high accuracy on an independent and identically distributed (i.i.d.) test set. However, models that are highly accurate on average can still consistently fail on rare and atypical examples (Hovy & Sgaard, 2015; Blodgett et al., 2016; Tatman, 2017; Hashimoto et al., 2018; Duchi et al., 2019). Such models are problematic when they violate equity considerations (Jurgens et al., 2017; Buolamwini & Gebru, 2018) or rely on *spurious correlations*: misleading heuristics that work for most training examples but do not always hold. For example, in natural language inference (NLI)—determining if two sentences agree or contradict—the presence of negation words like 'never' is strongly correlated with contradiction due to artifacts in crowdsourced training data (Gururangan et al., 2018; McCoy et al., 2019). A model that learns this spurious correlation would be accurate on average on an i.i.d. test set but suffer high error on groups of data where the correlation does not hold (e.g., the group of contradictory sentences with no negation words).

To avoid learning models that rely on spurious correlations and therefore suffer high loss on some groups of data, we instead train models to minimize the *worst-case* loss over groups in the training data. The choice of how to group the training data allows us to use our prior knowledge of spurious correlations, e.g., by grouping together contradictory sentences with no negation words in the NLI example above. This training procedure is an instance of distributionally robust optimization (DRO),

*Equal contribution.

1

Overview

- Introduction
  - Definitions and Motivation
- Paper Overviews
  - Fairness - Fair Attribute Classification through Latent Space De-biasing (Ramaswamy et al., 2021)
  - DRO - Distributionally Robust Neural Networks For Group Shifts (Sagawa et al., 2020)
- Fairness Experiment and Results
- DRO Experiment and Results
- Fairness + DRO Experiment and Results
- Conclusions

Problem Visualization

Image Regeneration with correlated features

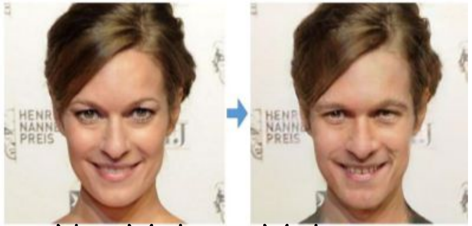
- Not Blond vs Male (Not Blond and Male)
- Most faces in training dataset with the attribute (“male” = 1) are also not blond.
    - Introduces (unintended) correlation between Not Blond and Male

Solution?

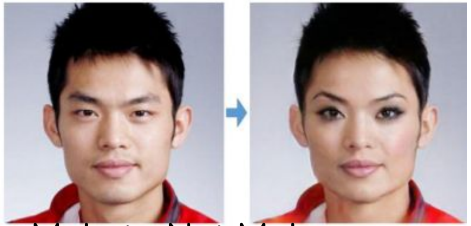
- Generalization
    - regularizers
    - preemptive analysis
- Add more data
    - Adversarial learning
    - Duplicating minority samples
    - synthetic data augmentation

Not Blond to Blond

Not Blond to Blond



Not Male to Male



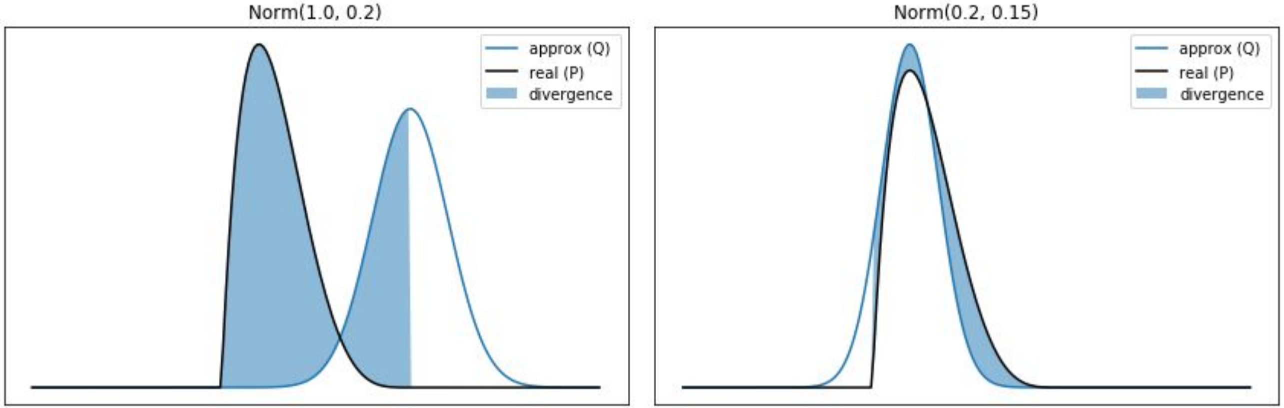
Male to Not Male

KL Divergence

$$D_{KL}(p||q) = \sum_{i=1}^{N} p(x_i) log(\frac{p(x_i)}{q(x_i)})$$

Kullback-Leibler (KL) Divergence: quantifies difference between probability distributions

Intuition: when the probability for an event in P is large, but the probability for the event in Q is small, there is a large divergence. When P and Q are reversed, it is not as large.



GAN-Debiasing

- Use Generative Adversarial Networks (GANs) to "even out" the dataset

For target label *t* and protected attribute *g*
Goal: separate target features from protected features.

Method:

1) Train classifiers $f_t(x)$ and $f_g(x)$ on original images $\mathcal{X}$ (CelebA)
2) Use a GAN trained on real images $\mathcal{X}$ whose generator G generates a synthetic image $\mathbf{x}$ from a random latent vector $\mathbf{z} \in \mathcal{Z}$
3) Assign semantic attribute labels using $f_t(x)$ and $f_g(x)$

The GAN naturally inherits biases from training data → *latent vector perturbation*

- Sample random set of latent vectors (with inherited biases)
- Train classifiers h_t, h_g in latent space that approximate $f_t \circ G, \quad f_g \circ G$
- Generate a complementary latent vector $\mathbf{z}'$ with same protected label but opposite protected label:

$$h_t(\mathbf{z}') = h_t(\mathbf{z}), \quad h_g(\mathbf{z}') = -h_g(\mathbf{z})$$

This creates a data generation method that is agnostic to classifier used to compute h

Model parameters:

- Epochs: 20
- Batch size: 32
- Optimizer: Adam
- Image size: 128x128
- Train/Test Split: 80:20
- Architecture: ResNet-50

GAN-Debiasing Results

Target Attribute: Blond

Protected Attribute: Male

Dataset	Training Method	Val/Test	Accuracy	KL-Divergence
Original CelebA Dataset	Linear SVM	Validation	0.9239	0.7125
		Test	0.9094	0.4603
Augmented CelebA Dataset		Validation	0.9976	0.1295
		Test	0.9954	0.2385

Time Complexity:

- On a Nvidia 1660Ti GPU 6 GB, ~38 min per epoch to train the classifier on the entire dataset ~202000 images.
- Generating the scores (predicted classifications) for each of the generated images: < 10 minutes
- Generating the Images: ~35 minutes (Generates 175,000 Images)

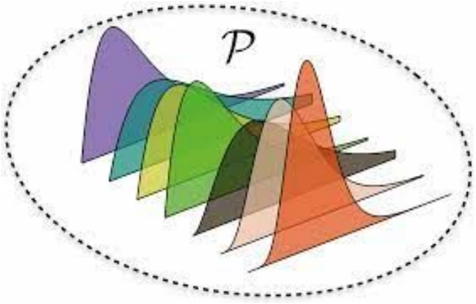
Overview

- Introduction
  - Motivation and Definitions
- Paper Overviews
  - Fairness - Fair Attribute Classification through Latent Space De-biasing (Ramaswamy et al., 2021)
  - DRO - Distributionally Robust Neural Networks For Group Shifts (Sagawa et al., 2020)
- Fairness Experiment and Results
- DRO Experiment and Results
- Fairness + DRO Experiment and Results
- Conclusions

Distributionally Robust Optimization (DRO)

Goal: minimize worst-case loss among “groups” in the dataset

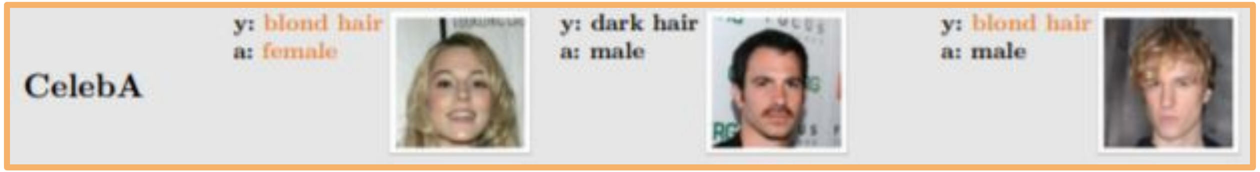
Models often learn “spurious” correlations between attributes, which causes them to perform poorly on certain subsets of their training data



Distributional Robustness on the CelebA Dataset

- 162770 training examples
  - 1387 in the smallest group (blond-haired males)
- Target attribute: {blond, not blond}
- Spurious attribute: {male, female}
- We use the following optimization problem:
  - ***q*** is a distribution over groups with high masses on high-loss groups
  - ***g*** represents a subgroup of the data, ***m*** denotes the number of subgroups (4 in our case)

$$\min_{\theta\in\Theta} \sup_{q\in\Delta_m} \sum_{g=1}^{m} q_g \mathbb{E}_{(x,y)\sim P_g}[\ell(\theta;(x,y))]$$



DRO Algorithm

Algorithm 1: Online optimization algorithm for group DRO

Input: Step sizes η_q, η_θ; P_g for each $g \in \mathcal{G}$

Initialize $\theta^{(0)}$ and $q^{(0)}$

for $t = 1, \dots, T$ **do**

- $g \sim \text{Uniform}(1, \dots, m)$ `// Choose a group` g `at random`
- $x, y \sim P_g$ `// Sample` x, y `from group` g
- $q' \leftarrow q^{(t-1)}; \ q'_g \leftarrow q'_g \exp(\eta_q \ell(\theta^{(t-1)}; (x, y)))$ `// Update weights for group` g
- $q^{(t)} \leftarrow q' / \sum_{g'} q'_{g'}$ `// Renormalize` q
- $\theta^{(t)} \leftarrow \theta^{(t-1)} - \eta_\theta q_g^{(t)} \nabla \ell(\theta^{(t-1)}; (x, y))$ `// Use` q `to update` θ

end

The general idea of the algorithm[2] is to allow groups with high loss (usually underrepresented groups) have a greater impact in adjusting the weights of the model

[2] Sagawa et al. (2020). Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization.

DRO Algorithm

$q_1 = 0.187$

Label: not blond
Spurious attribute: female

n_1=2826

$q_2 = 0.093$

Label: not blond
Spurious attribute: male

n_2=4753

$q_3 = 0.209$

Label: blond
Spurious attribute: female

n_3=2274

$q_4 = 0.511$

Label: blond
Spurious attribute: male

n_4=147

DRO Experiment Setup with CelebA Dataset

- Binary classification model - ResNet-50
  - Input: Images normalized to a specified size (i.e. 128x128)
  - Output: {0,1} - prediction of {not blond, blond}
- Data
  - Subset of CelebA dataset: 10000 randomly sampled images (with 80/20 train/test split)
- Metrics
  - Overall accuracy
  - Accuracy across each group: ({blond, male}, {blond, female}, {not blond, male}, {not blond, female}) - specifically keep track of worst-case accuracy
- Training parameters
  - Parameters of interest
    - ERM vs. DRO training
    - L2-penalty (0.0001 or 0.01)
  - Constant parameters
    - 20 epochs
    - SGD optimizer
    - Batch size = 4

DRO Experiment Results

Training Method	Accuracy	Worst-Group Accuracy	KL-Divergence
ERM w/ standard reg.	0.7928	0.0612	0.0890
ERM w/ strong L2 penalty	0.8218	0.0408	0.1521
DRO w/ standard reg.	0.7688	0.1224	0.0739
DRO w/ strong L2 penalty	0.8596	0.0000	1.2370

- Cost of algorithm (On a Nvidia 1660Ti GPU 6 GB):
  - ERM- 10 minutes per epoch
  - DRO- 15 minutes per epoch
- DRO with standard regularization had the best worst-group accuracy and KL-divergence but decreased overall accuracy

Overview

Combining the Fairness and DRO Approaches

Method:

- Train our model using the DRO algorithm on the augmented dataset

Constraints:

- Memory: DRO needs to parameters learned for each group, which grows linearly with each epoch. So our computers and Colab ran out of memory.
- Time: The models take 10~15 minutes per epochs thus restricting how long we were able to train each model.

Expectations:

- We see that many of the augmented images are noisy or distorted, we expect DRO to be able to remove the noise and use the generated features to create a better classifier.

Fairness + DRO Experiment Results

	Training Method	Accuracy	Worst-Group Accuracy	KL-Divergence
Original CelebA Dataset	ERM w/ standard reg.	0.7928	0.0612	0.0890
	ERM w/ strong L2 penalty	0.8218	0.0408	0.1521
	DRO w/ standard reg.	0.7688	0.1224	0.0739
	DRO w/ strong L2 penalty	0.8596	0.0000	1.2370

	Training Method	Accuracy	Worst-Group Accuracy	KL-Divergence
Augmented CelebA Dataset	ERM w/ standard reg.	0.8026	0.0816	0.1061
	ERM w/ strong L2 penalty	0.8536	0.0046	0.5560
	DRO w/ standard reg.	0.8598	0.0000	1.2348
	DRO w/ strong L2 penalty	0.8596	0.0000	1.2370

Time Complexity:

ERM: ~10 min/epoch

DRO: ~15 min/epoch

Observation:

- Appears to be a tradeoff between fairness (KL Divergence) and distributional robustness (worst-group accuracy)

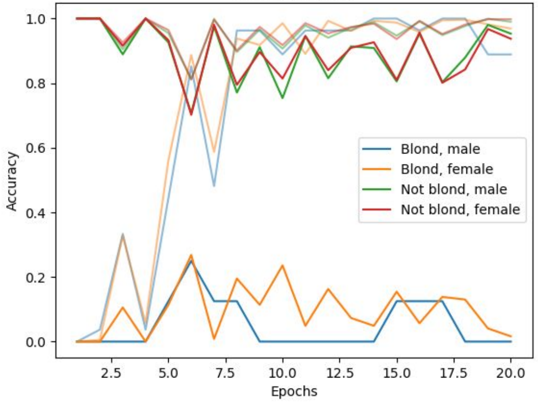
Overview

Conclusions

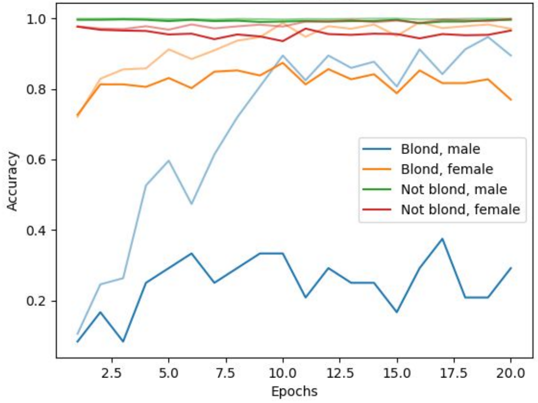
Expectations:

- DRO will outperform ERM worst-group accuracy, will be similar in KL Divergence, but will be slightly worse in overall accuracy
- With augmented data, DRO will outperform ERM in worst-group accuracy and KL divergence, but will underperform in overall accuracy.

ERM



DRO



*for the plots: dark lines indicate test accuracy, translucent lines indicate training accuracy

Key Takeaways

Tradeoff between Fairness Augmentation and DRO

- Using a fairness augmented dataset decreased the accuracy, worst-group accuracy, and KL Divergence.

However there was significant hardware limits, so by using a larger GPU or a computer with more RAM, we could train for larger on larger subsets to train a better model. (AWS plz return my calls)

References

[1] Vikram V. Ramaswamy, Sunnie S. Y. Kim, & Olga Russakovsky. (2021). Fair Attribute Classification through Latent Space De-biasing.

[2] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, & Percy Liang. (2020). Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization.

[3] Ziwei Liu, Ping Luo, Xiaogang Wang, & Xiaoou Tang. (2015). Deep Learning Face Attributes in the Wild.