Fair Image Classification with Semi-Supervised Learning

Caroline Ho
Department of Computer Science
Stanford University
Stanford, CA 94305
cho19@stanford.edu

Hugo Kitano
Department of Computer Science
Stanford University
Stanford, CA 94305
hkitano@stanford.edu

Kevin Lee
Department of Electrical Engineering
Stanford University
Stanford, CA 94305
kelelee@stanford.edu

Abstract

In modern computer vision, which is powered by the availability of large amounts of data, labeling each data point with a protected class label in order to train a fair model can be prohibitively costly. Hence, our goal is to develop a fair image classifier for datasets in which only a subset of images is labeled with the protected class. We found that leveraging adversarial learning, we were able to improve fairness metrics with minimal decrease in accuracy.

1 Introduction

The accumulation of large amounts of data with protected class labels is a major barrier to the training of fair deep learning models. In computer vision, this problem can be overwhelming: the brute force method of labeling classes by hand is simply unfeasible. This can be compounded by the fact that protected classes are often minorities and less numerous compared to the general population. Our project will attempt to make fair predictions on an outcome variable Y with only a subset of the data labeled with the protected class A. This model architecture could be a powerful debiasing tool that can be used for many applications in computer vision that aim to make machine learning fairer and more accountable, as well as a useful method for data augmentation when the cost of acquiring images with protected class labels is expensive.

2 Data

We use the CelebA dataset [1], which contains over 200,000 images of celebrities with 40 binary attribute annotations (We include a visualization in Figure 2). The dataset is split in into train, validation, and test sets by its creators. Each image has been cropped into 178×218 pixel images and center-aligned around the face. We define the gender attribute as a protected class *A* (which we do not include in the baseline model inputs). Our task is to predict the 39 other attributes, *Y*, from each image. Many of these attributes are very gender-imbalanced, as shown in Figure 1.

As a pre-processing step in our experiments, we transform each image to 224×224 pixels and normalize by ImageNet mean and standard deviation.

5_o_Clock_Shadow	0.9991	**Chubby**	0.8765	**Pointy_Nose**	0.2439
Arched_Eyebrows	0.0834	**Double_Chin**	0.8790	**Receding_Hairline**	0.6133
Attractive	0.2270	**Eyeglasses**	0.7942	**Rosy_Cheeks**	0.0190
Bags_Under_Eyes	0.7094	**Goatee**	0.9989	**Sideburns**	0.9990
Bald	0.9962	**Gray_Hair**	0.8512	**Smiling**	0.3460
Bangs	0.2263	**Heavy_Makeup**	0.0029	**Straight_Hair**	0.4848
Big_Lips	0.2701	**High_Cheekbones**	0.2817	**Wavy_Hair**	0.1836
Big_Nose	0.7456	**Mouth_Slightly_Open**	0.3659	**Wearing_Earrings**	0.0352
Black_Hair	0.5189	**Mustache**	0.9996	**Wearing_Hat**	0.6998
Blond_Hair	0.0583	**Narrow_Eyes**	0.4355	**Wearing_Lipstick**	0.0054
Blurry	0.4686	**No_Beard**	0.3022	**Wearing_Necklace**	0.0604
Brown_Hair	0.3076	**Oval_Face**	0.3225	**Wearing_Necktie**	0.9976
Bushy_Eyebrows	0.7159	**Pale_Skin**	0.2362	**Young**	0.3410

Figure 1: Percentage of positive-labeled datapoints for each attribute that are male

Figure 2: Samples from the CelebA dataset grouped by attributes.

3 Approach

3.1 Background

We follow an earlier work by Beutel et al. [2] that uses a model that tries to predict the target class Y, while simultaneously preventing an adversarial head from predicting the protected class A from its latent representation of the data. Only a subset of the training data is labeled with protected class information. As training progresses, the encoded representation of the datapoint evolves so that the adversarial head cannot predict its protected class label, thus becoming an unbiased representation of the data. Notably, the paper produces poor results if the labeled subset of the data is skewed, so in our semi-supervised experiments, we ensure that the subset is balanced.

3.2 Model architecture

Beutel et al.'s experiments were done on tabular data, so we make several changes to adapt their method to handle image data. As with their model, our model is composed of three parts: an encoder, a classifier (primary head), and an adversarial head. We include our adapted model architecture in Figure 3 and provide a detailed description below.

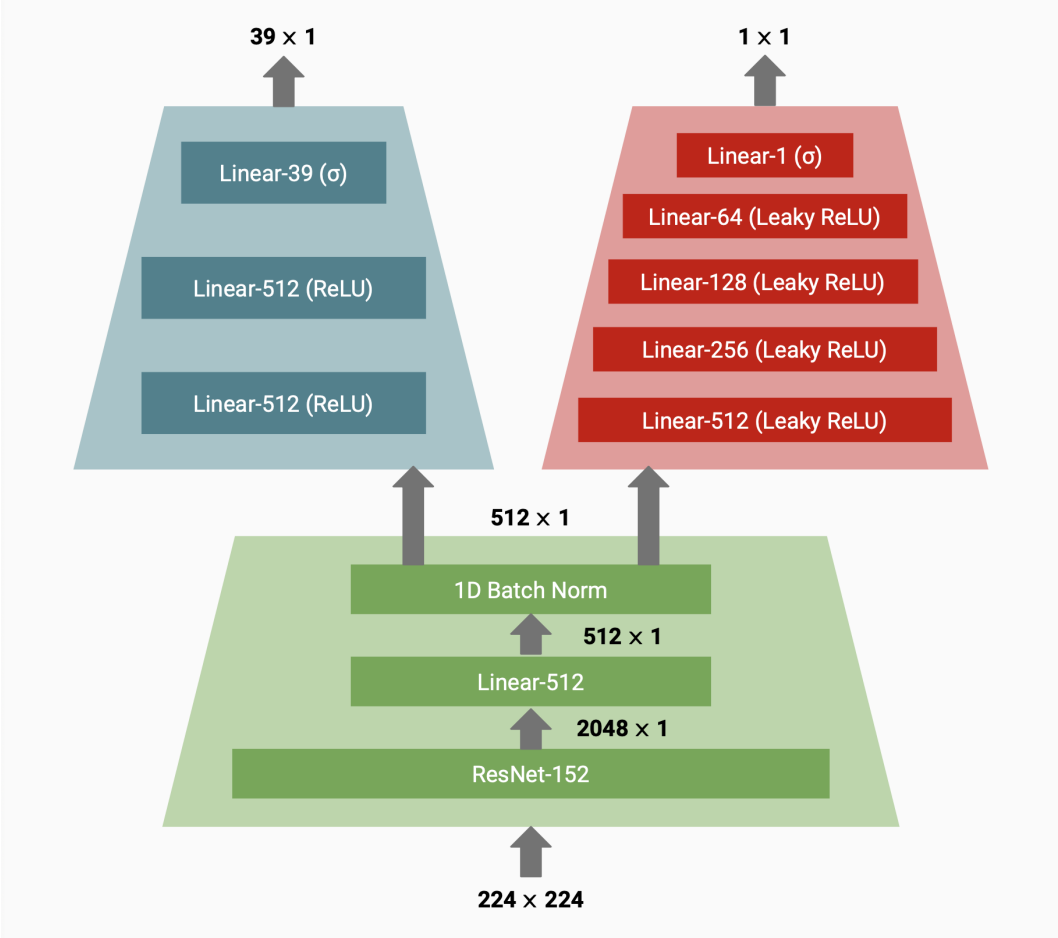

Figure 3: Model architecture (adapted from Beutel et al.). Green, cyan and red blocks are the shared encoder, primary head and adversarial head, respectively.

Encoder. Our encoder is a ResNet-152 pre-trained on ImageNet, with the final fully connected layer replaced with (1) a linear layer with an output size of `hidden_size` $= 512$ and (2) a one-dimensional batch normalization layer.

Classifier. Our classifier contains (1) two ReLU-activated linear layers with output sizes of 512 and (2) a final linear layer with an output size of 39, thus outputting logits for all 39 attributes.

Adversarial head. The adversarial head takes in only the latent representation of datapoints that have protected class labels and outputs the probability of the datapoint belonging to the protected class. The adversarial head is made up of five linear layers of decreasing output size (512, 256, 128, 64, and 1 neurons, in order) and uses the Leaky-ReLU activation function for the first four layers.

3.3 Training procedure

During training, all images are fed through the encoder and classifier. For images that have protected class labels, their latent representations (their encoder outputs) are given to the adversarial head. Our adversarial head is trained with its own loss and optimization, while the encoder and classifier are trained with the following composite loss function (where λ is a user-specified hyperparameter representing de-biasing strength):

$$L_{primary} = L_{classification} - \lambda * L_{adversarial} \tag{1}$$

During validation and testing, the adversarial head is not touched, and only the encoder and classifier are used.

3.4 Baseline

In line with Beutel et al. [2], our baseline model only comprises an encoder module that feeds into a classifier module, omitting the adversarial head (and thus not performing de-biasing).

4 Experiments

Our code can be found at `https://github.com/carolineh101/debiasing-images`.

We implement all of our models and experiments ourselves using PyTorch [3]. We run several experiments using the following parameters:

- $\lambda \in \{0.2, 0.4, 0.6, 0.8, 1.0, 2.0\}$
- $\phi \in \{0.00, 0.01, 0.05, 0.10, 0.25, 1.00\}$

Here, ϕ is the fraction of training data labeled with protected attribute information. $\phi = 0.00$ represents the baseline, and all other values of ϕ represent subsets balanced by gender except for $\phi = 1$ (for which all data points are labeled). The balanced subsets are set up with exactly half of the data belonging to the protected class.

4.1 Evaluation method

We evaluate our results using accuracy as well as two fairness metrics: parity gap (which helps us understand demographic parity) and equality gap (which helps us understand equality of opportunity). Our goal is to maximize accuracy while minimizing the fairness metrics (bounded between 0 and 1).

Beutel et al. [2] define the fairness metrics as follows:

$$ParityGap = |ProbTrue_1 - ProbTrue_0| \tag{2}$$

$$EqualityGap_y = |ProbCorrect_{y,1} - ProbCorrect_{y,0}| \tag{3}$$

In our experiments, we deal with two equality gaps: one for the positive Y label $Y = 1$, and another for the negative Y label $Y = 0$. Beutel et al. further define $ProbTrue_a$ and $ProbCorrect_{y,a}$ as follows (where N_a is the number of examples in protected class a):

$$ProbTrue_a = P(\hat{Y} = 1|A = a) = \frac{TP_a + FP_a}{N_a} \tag{4}$$

$$ProbCorrect_{1,a} = P(\hat{Y} = 1|A = a, Y = 1) = \frac{TP_a}{TP_a + FN_a} \tag{5}$$

$$ProbCorrect_{0,a} = P(\hat{Y} = 0|A = a, Y = 0) = \frac{TN_a}{TN_a + FP_a} \tag{6}$$

4.2 Optimization and hyperparameters

We use binary cross-entropy loss for our primary and adversarial loss criteria (implemented with `BCEWithLogitsLoss`):

$$\ell(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \cdot \log \hat{y_i} + (1 - y_i) \cdot \log(1 - \hat{y_i})] \tag{7}$$

For our optimizer, we use Adam [4] with $B_1 = 0.9$ and $B_2 = 0.999$. When training, we use a batch size of 32, a primary learning rate of 0.00001, and an adversarial head learning rate of 0.0001. We run for 10 epochs and evaluate the model with the highest validation accuracy.

4.3 Results

We include aggregate results across all attributes in Figure 4, select highly gender-imbalanced per-attribute results in Figures 5, 6, and 7, and select fairly gender-balanced per-attribute results in Figures 8 and 9.

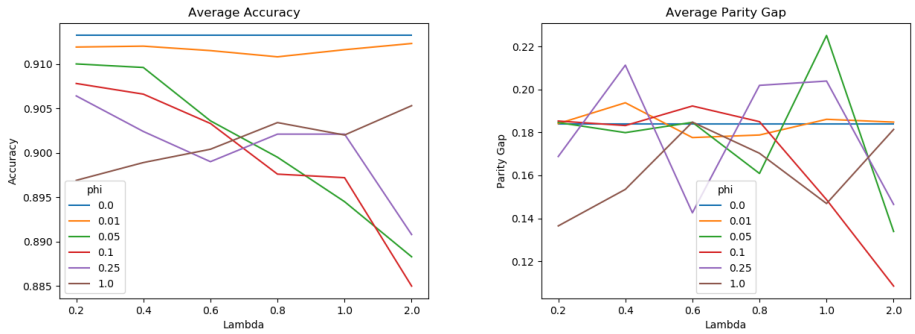

Figure 4: Average performance and fairness results.

4.3.1 Average results

In Figure 4, we can see decent reduction in parity gap averaged over the attributes, while decrease in accuracy is minimal. Generally, as λ increases, the model learned is fairer, which makes sense since increasing λ increases the strength of the adversarial head. The increase of ϕ is also generally correlated with greater fairness, though the performance of $\phi = 1$ seems to degrade as λ increases. We suspect that this is because when we label all points in our dataset with the protected class, the data is no longer balanced by protected class, which weakens the strength of the adversarial head. Thus, we recommend balancing the subset of data with protected class labels in all future experiments.

4.3.2 Gender-imbalanced attribute results

We now look specifically at three highly gender-imbalanced hand-picked attributes. Positive examples of the attribute `Blond_Hair` are only 5.83% male, which seems to us to be artificially low. We are able to show in Figure 5 that fairness training lowers the parity gap and equality gap for positive labels (blond hair) quite substantially, with extremely minimal decrease in accuracy.

In Figure 6, we see the `Rosy_Cheeks` attribute trains less stably than `Blond_Hair`. Only for $\lambda = 2$ does the fairness training inarguably lead to fairer results.

Results for `Wearing_Necktie` in Figure 7 similarly show a small decrease in accuracy and improvement in fairness across most adversarial settings (with the exception of the $\lambda = 1.0$, $\phi = 0.05$ setting, which performs very poorly on both accuracy and fairness metrics). This type of random result suggests that multiple models should be trained in order to avoid such uncommonly poor models.

4.3.3 Gender-balanced attribute results

Examining the performance of our model on two fairly gender-balanced attributes, namely `Black_Hair` (Figure 8) and `Straight_Hair` (Figure 9), it seems that accuracy decreases somewhat from the baseline for most adversarial settings, as with the gender-imbalanced case. However, fairness results seem far less stable and much poorer in performance, with most settings actually leading to increases in the parity and equality gaps from the baseline. Thus, most of the increased fairness from adversarial learning is the result of increased fairness from heavily imbalanced attributes.

5 Conclusion

Our experiments demonstrate that training with an adversarial head can increase fairness with minimal decrease in accuracy for attributes highly imbalanced in the protected class (adversarial training does not appear to be worth the trade-off for the gender-balanced case). Although having more data labeled with the protected class does make the model fairer, the model can still become fairer with the use of only a small percentage of data with protected class labels, especially when λ is large (as long as this percentage is balanced over the protected class).

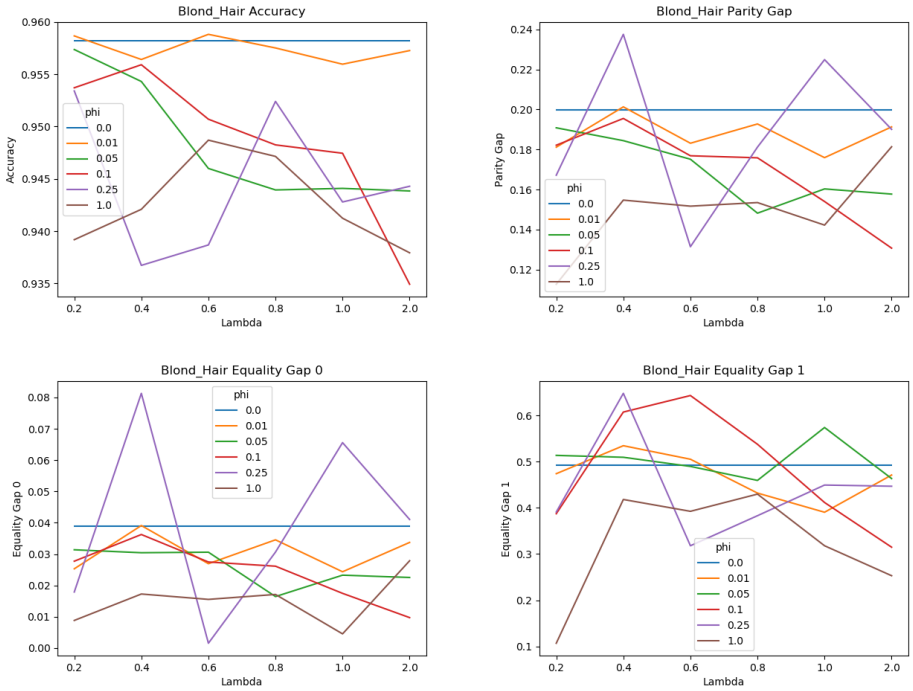

Figure 5: `Blond_Hair` (5.83% male) performance and fairness results.

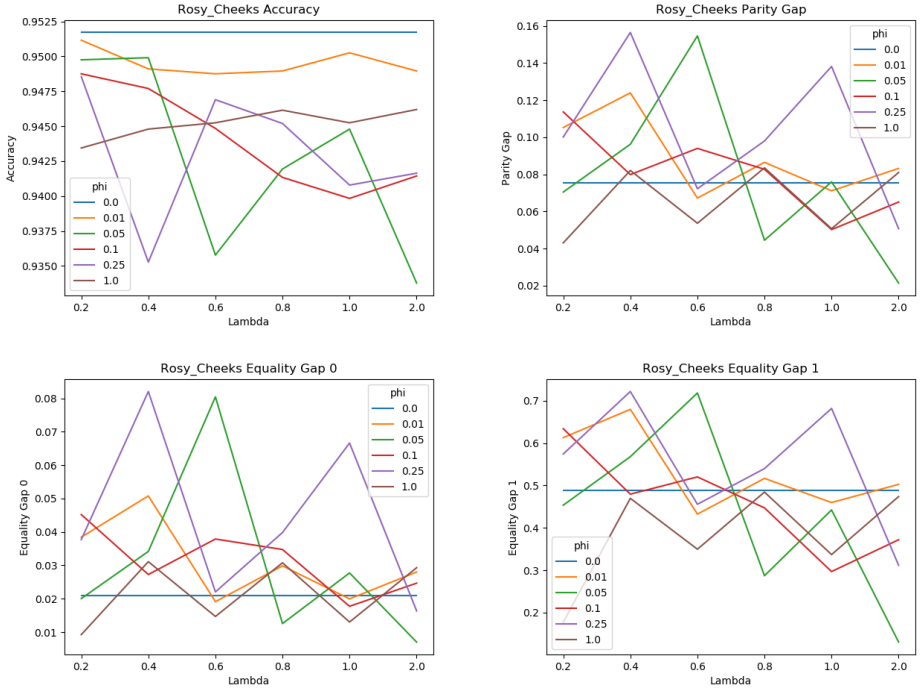

Figure 6: `Rosy_Cheeks` (1.90% male) performance and fairness results.

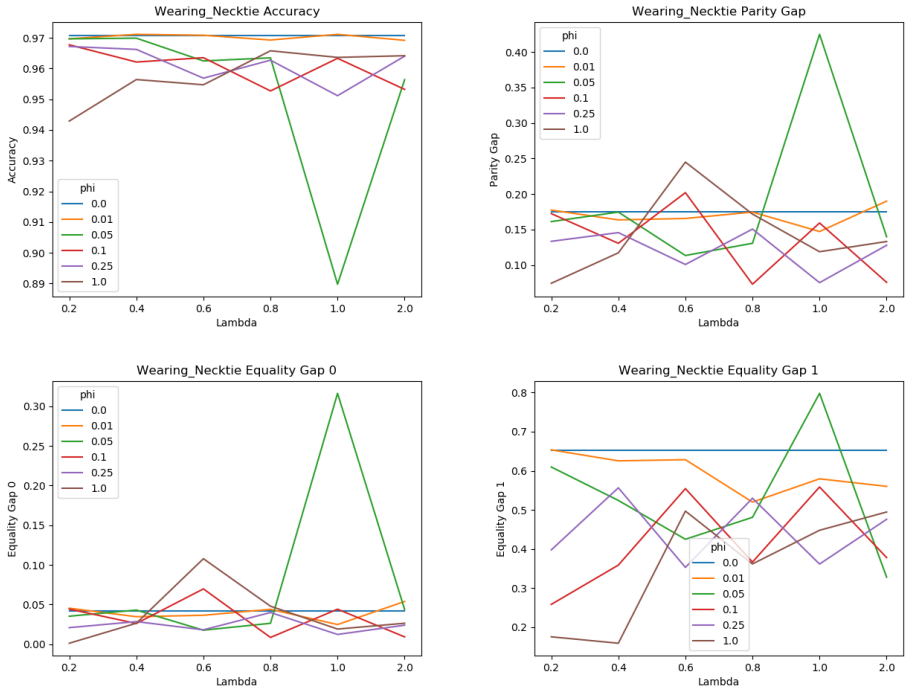

Figure 7: `Wearing_Necktie` (99.76% male) performance and fairness results.

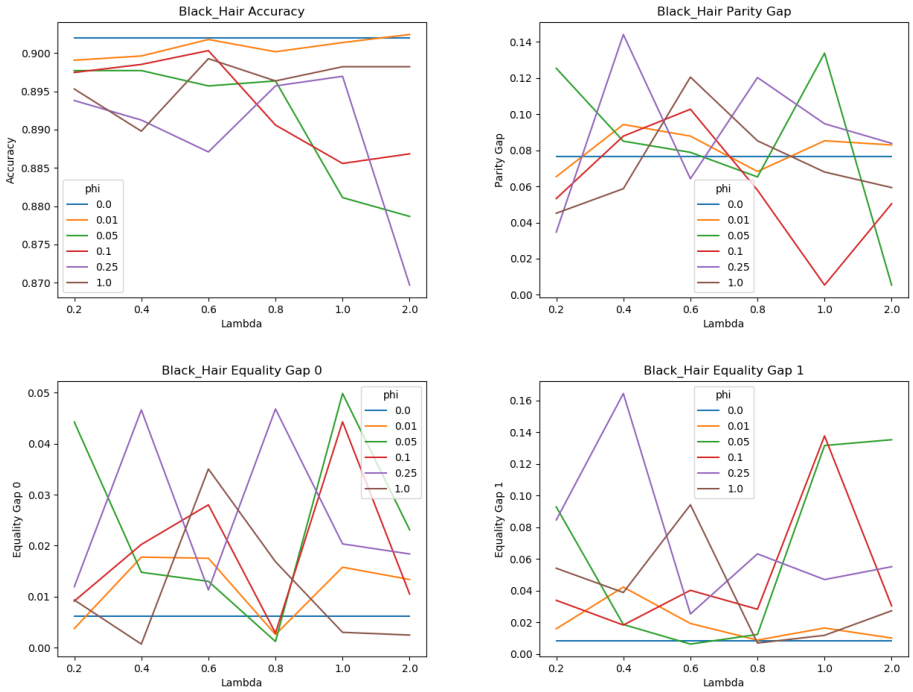

Figure 8: `Black_Hair` (51.89% male) performance and fairness results.

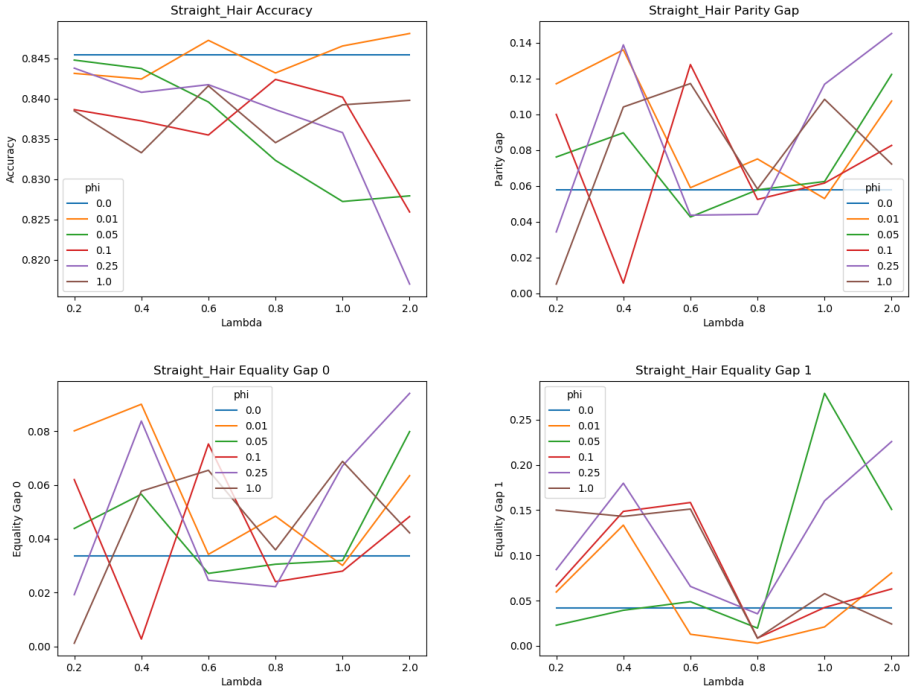

Figure 9: `Straight_Hair` (48.48% male) performance and fairness results.

5.1 Future work

Due to time and resource constraints, we did not explore all of the original approaches we considered. In the future, more extreme values of λ can be experimented with, for example, $\lambda = 5$ and $\lambda = 10$. We may see even smaller a gap in equality metrics at the cost of a drop in accuracy in these circumstances. Since Beutel et al. only tried values from 0 to 2, we thought that was sufficient for our project. We would also like to see the experiment replicated on another dataset that is less stratified by the protected class label, as we saw from CelebA.

Some additional avenues for future work might include exploring potential ways to adapt other methods for fair image classification, such as (1) Wang et al.'s Domain Independent Training method [5] and (2) Madras et al. and Xu et al.'s use of GANs [6] [7], for the semi-supervised case.

References

[1] Z. Liu, P. Luo, X. Wang, , and X. Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[2] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. Data decisions and theoretical implications when adversarially learning fair representations, 2017.

[3] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[4] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[5] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation. *CVPR*, 2019.

[6] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning Adversarially Fair and Transferable Representations, 2018.

[7] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. FairGAN: Fairness-aware Generative Adversarial Networks, 2018.