Reducing Bias in a Facial Gender and Age Predictor

Jack Irish
Stanford University
jirish42@stanford.edu

Dhruv Tandon
Stanford University
dtandon@stanford.edu

Abstract

Estimating demographic information from images of human faces is a popular and well-studied application of deep learning techniques. However, as these deep networks see more real-world use, it is becoming evident that their accuracy is biased towards certain demographics. Leveraging a transfer learning approach with the EfficientNetV2 architecture, our project evaluates the performance of various bias reduction techniques, including variance penalization, loss scaling, hierarchical predictors, and data augmentation. Using the UTKFace dataset, our methods demonstrated improvements in fairness for age estimation across different racial groups, albeit with some trade-offs in overall accuracy. The direct variance penalization method was notably effective, achieving a balanced improvement in fairness and modest accuracy reduction. Future work aims to expand upon these findings by employing larger, more diverse datasets and refining augmentation techniques to further enhance model performance and fairness.

1. Introduction

In recent years, deep learning, particularly through the use of Convolutional Neural Networks (CNNs), has revolutionized facial recognition technology, enabling significant advancements across many frontiers. Although this has proven to be extremely valuable in security systems, marketing, and the search for missing persons or criminal suspects, such critical areas demands a high level of accuracy and consistency across subjects of different racial backgrounds. Racial bias in facial detection algorithms remains a significant challenge and can have disastrous consequences, especially when these models are employed in security or policing applications [1]. One particular instance of this occurred in 2018 when the American Civil Liberties Union (ACLU) tested Amazon's Rekognition software by running a facial recognition scan of members of Congress against a database of mugshots. The software incorrectly matched 28 members, disproportionately identifying people of color as criminals [2]. Project Green Light, another instance of racial bias in facial recognition technology, involved the installation of high-definition surveillance cameras in predominantly Black neighborhoods in Detroit. These cameras were linked to real-time facial recognition systems, resulting in increased surveillance and racial profiling of Black residents [3].

2. Related Work

Recent years have seen a number of proposed systems for estimation of demographic traits from facial images. In 2012, Karimi et al. [4] developed a method that used classical computer vision techniques to extract the location of facial features and inferred the subject's gender and age from ratios between the feature locations. Later in the same decade, the advent of deep convolutional neural networks and the growing availability of accelerated hardware to train on led to a multitude of deep learning based approaches to the gender and age prediction problem [5][6][7]. In 2020, Abdolrashidi et al. [8] designed an ensemble model consisting of Residual Attentional and ResNet architectures. The authors also proposed a technique in which the result of the gender classifier is fed into the age predictor in order to improve its estimation. The next year, Garain et al. [9] published an improved Gated Residual Attentional architecture for the same gender and age problem. This model achieved impressive age estimation accuracy by splitting the regression problem into the classification of the decade and regression of the remainder (age % 10). However, novel deep network architectures are not always necessary to achieve good performance. Smith and Chen [10] demonstrated that transfer learning produces competitive age and gender prediction accuracies, even when using a general feature extracting model that is not pretrained on facial images. Finally, although bias and fairness are just beginning to gain attention in the field of deep learning, work has been done to formalize the problem and begin to develop solutions [11][12]. In a 2021 paper, Feldman and Peake [13] defined several metrics for fairness in deep learning and compared the results of methods to reduce gender bias in a model de-

signed to predict adult income level.

3. Methods

As our research is more concerned with fairness than absolute performance, we chose to use a transfer learning-based architecture similar to the models proposed by Smith and Chen [10]. This approach had a number of advantages, given the limited time and resources available for this project. The use of a pre-trained deep network in our models allowed us to leverage state of the art feature representations without the immense engineering effort required to design a sophisticated convolutional network. Further, fine-tuning smaller models on top of this larger base model greatly accelerated our design and training process due to the fact that only a fraction of our model's parameters required optimization. Transfer learning was also an appropriate methodology for our research due to the relatively small size of our dataset. In order to appropriately evaluate our model's performance among different races, we needed images that were accurately labeled with age, gender and race, which are not publicly available in numbers much more than a few tens of thousands. A general purpose feature extractor trained on a large-scale dataset allowed us to approach state-of-the-art performance without the need to manually collect any additional data.

We used the EfficientNetV2 model, first proposed by Tan and Le [14] and trained on ImageNet, as our pre-trained feature extractor. We chose to use the Medium variant of the model, with 7 convolutional blocks and 54.1 million parameters. The EfficientNetV2 architecture has been shown to outperform many state-of-the-art networks on ImageNet classification as well as transfer learning tasks while boasting significantly reduced training time and parameter count. To achieve this, the authors used novel, data-driven techniques to search the design space for an optimal arrangement of two main convolutional blocks (Figure 1). EfficientNetV2-M takes 128 x 128 to 380 x 380 RGB images as input and returns scores over the 1000 ImageNet classes. However, we are only using the layers up to the first fully connected classifying layer and fine-tuning our own prediction heads, so the pre-trained portion of EfficientNetV2-M that we used for the following models effectively produces a 1280-dimensional feature vector as an output. We used PyTorch [15], Torchvision [16], Matplotlib [17], and NumPy [18] libraries throughout the project.

3.1. Baseline Model

In order to identify biases and establish a baseline against which we could gauge the fairness of our proposed methods, we designed a standard dual-objective gender and age prediction model based on the work of Smith and Chen [10]. The baseline model consists of separate fully-connected heads to perform gender classification and age regression.

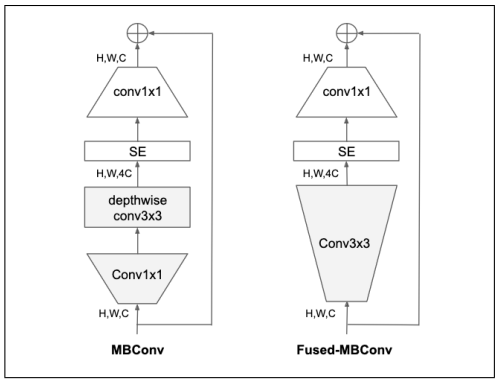

Figure 1. EfficientNetV2 Convolutional Blocks

Both heads take a 1280-vector of features from the pre-trained EfficientNetV2 model as input. The gender classification head produces two class scores for Male and Female while the age regressor produces a single output, the estimated age. The gender head consists of two hidden layers with 256 neurons each, followed by an output layer of size 2 (Figure 2). The age head has two hidden layers of size 2048 followed by an output layer of size 1 (Figure 3). For both heads, each hidden layer uses ReLU activation and is followed by a dropout layer in order to regularize the model.

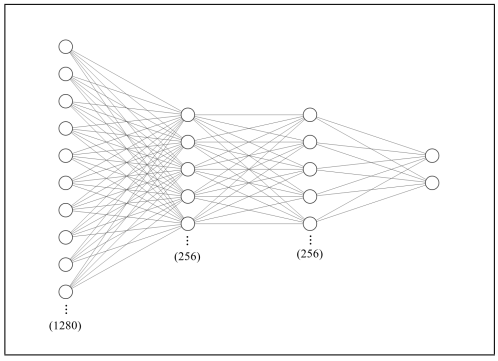

Figure 2. Baseline Gender Head

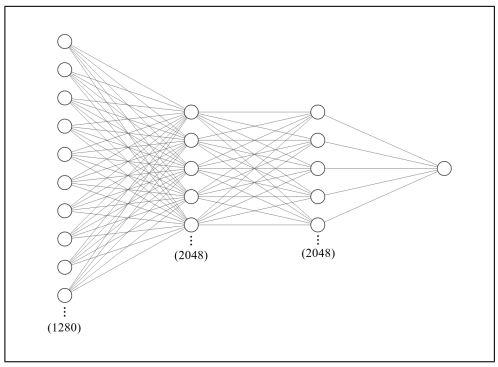

Figure 3. Baseline Age Head

The two heads were trained independently, using cross-entropy loss (Equation 1) for the gender classifier and L1

loss (Equation 2) for the age regressor. Cross-entropy is the natural choice for classification loss, but we specifically chose to use L1 loss for the age regressor due to its improved robustness to outliers over methods like MSE loss (Equation 3) and because Mean Average Error (MAE), an identical statistic, is commonly used to evaluate the accuracy of such an age estimator [9][10].

$$\frac{1}{N}\sum_{i=1}^{N} -\log\frac{\exp(x_{i,y_i})}{\sum_{c=1}^{C}\exp(x_{i,c})} \tag{1}$$

$$\frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \tag{2}$$

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{3}$$

3.2. Bias Reduction Methods

Next, we outline four different bias reduction methods designed to equalize the model's performance over the dataset's five race categories.

3.2.1 Direct Variance Penalization

For our first bias reduction method, we added a term to the original L1 or MAE loss to directly penalize high variance in performance over the five protected race classes. This is straightforward for the age regression task, since the variance of mean average errors is differentiable with respect to model outputs (Equation 4). However, because classification accuracy is not differentiable, we instead computed the variance of each race's cross entropy loss. (Equation 5). In both losses, v is a hyperparameter used to scale the impact of the variance term on the overall loss.

$$\begin{aligned}\hat{L} &= \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| + v\cdot\frac{\sum_{r=0}^{4}(P_r - \overline{P})^2}{5},\\ P_r &= \frac{\sum_{i=1}^{N}|\hat{y}_i - y_i|\cdot\mathbf{1}(\text{race}_i = r)}{\sum_{i=1}^{N}\mathbf{1}(\text{race}_i = r)}\end{aligned} \tag{4}$$

$$\begin{aligned}\hat{L} &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(x_{i,y_i})}{\sum_{c=1}^{C}\exp(x_{i,c})} + v\cdot\frac{\sum_{r=0}^{4}(P_r - \overline{P})^2}{5},\\ P_r &= \frac{-\mathbf{1}(\text{race}_i = r)\cdot\log\frac{\exp(x_{i,y_i})}{\sum_{c=1}^{C}\exp(x_{i,c})}}{\sum_{i=1}^{N}\mathbf{1}(\text{race}_i = r)}\end{aligned} \tag{5}$$

3.2.2 Loss Scaling

For our next method, we weigh the average loss over each protected race based on predetermined scalars, α_r, derived from the baseline model's performance for that sample's race. Each race's scaling factor is inversely proportional to its baseline performance (races for which the baseline model performs poorly are given more weight in the adjusted loss and vice versa). The scaling factors are normalized to the range $(0, 1)$ then passed through the softmax function with a moderation hyperparameter, m, used to control the difference in magnitude of scalars for different races (Equation 6). Note that here, baseline performance, P_b, refers to the baseline MAE for the age regressor and the *inverse* of classification accuracy for the gender classifier, since the scaling factors should be larger for races that perform worse. The final calculation of adjusted loss is shown in Equation 7, with L_i representing the original sample loss, either cross-entropy or L1.

$$\alpha_r = \text{softmax}(m\cdot\frac{P_{b,r} - \min_r P_{b,r}}{\max_r(P_{b,r} - \min_r P_{b,r})}) \tag{6}$$

$$\hat{L} = \sum_{r=0}^{4}\alpha_r\cdot\frac{\sum_{i=1}^{N}L_i\cdot\mathbf{1}(\text{race}_i = r)}{\sum_{i=1}^{N}\mathbf{1}(\text{race}_i = r)} \tag{7}$$

3.2.3 Hierarchical Predictor

For the next method, we train individual models for each race such that each model achieves roughly equal validation accuracy on its respective race. This is accomplished by adjusting training parameters like learning rate to produce models with the same performance as the worst-performing race from the baseline model. Then, a separate race classification model is trained and used to estimate a sample image's protected class and pass the sample through the appropriate race's model (Figure 4). In our experiments, we used a fully-connected race classifier with two hidden layers of size 512, ReLU activation and dropout, and trained with cross-entropy loss. At test time, if the race classifier's prediction does not have a confidence score above a certain threshold, the model will predict the "Others" class, even if the "Others" class did not have the highest score. We found that this adjustment improved the race predictor's performance, and thereby improved the performance of the hierarchical model.

3.2.4 Data Augmentation

Data augmentation is an very common and important technique in training deep neural networks for image recognition tasks. We decided so implement this in our code by

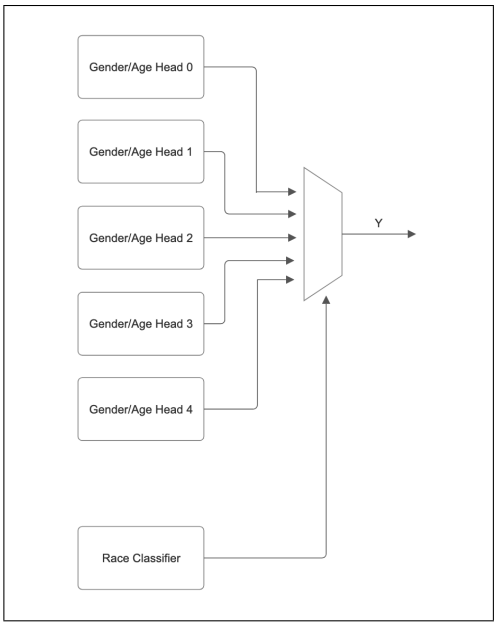

Figure 4. Hierarchical Predictor Architecture

creating new training samples by applying various transformations to the existing data based on the baseline results. The main factor we had to keep in mind was the fact that any augmentations must preserve the integrity of facial features, e.g. avoiding color jitters transformations as it could distort skin tones. This helps in creating a more equitable model that performs consistently across different racial groups and mitigate any inherent bias in our model. Similarly there are several transformations like random brightness adjustment, slight rotations, blurs and distortions can mimic real world variations making the model more accurate in recognizing faces even under sub optimal conditions.

Once we load the training dataset, we calculate the current distribution of races say $\{n_0, n_1, n_2, n_3, n_4\}$. Given baseline loss for each race r, represented as L_r, we apply an exponential function to introduce non-linearity making the us more sensitive to worse performing classes. The normalized exponential adjustment for race is given as:

$$\hat{e}_r = \frac{\exp\left(\frac{L_r}{\alpha}\right)}{\sum_{j=0}^{4} \exp\left(\frac{L_r}{\alpha}\right)} \tag{8}$$

The target augmented dataset is then calculated as:

$$T_r = \max\left(\lfloor \hat{e}_r \cdot \nu \cdot N \rfloor, n_r\right) \tag{9}$$

, where α is moderation factor that can be fine tuned in later stages, N is the total number of points in the original training distribution and ν is distribution scaling that set the size of target dataset as multiple of original training dataset. This method adjusts the representation of each class based on baseline losses, ensuring that under performing classes are given the necessary attention during the augmentation process.

4. Datasets

UTKFace dataset contains approximately 24,000 well-cropped and aligned human faces, annotated with key demographic attributes including age, gender, and ethnicity. This data set is suitable for tasks such as age estimation, gender classification, and facial detection. Specifically, ages range from 0 to 116 years, providing a broad spectrum for age estimation tasks, but we only have binary values representing the gender of the subject. The ethnicity label has a categorical value representing the ethnic background of the subject, with categories including White, Black, Asian, Indian, and others.

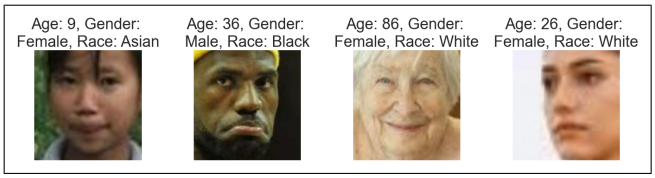

Figure 5. Training set image examples with the age, gender, and ethnicity labels

The dataset distribution across each labels is shown in Figure 6. Age histogram shows a a significant variation across different age groups with the mean around 33 years. The dataset is relatively balanced in terms of gender, with 52.2% of the images labeled as male and 47.8% labeled as female which is crucial for training a model that does not exhibit gender bias. Race distribution is critical for our case as we want to evaluate the model's performance across different racial groups and identify and mitigate any underlying biases. For our experiments, we used an 80%/10%/10% train/validate/test split, leading to just under 20,000 total training samples.

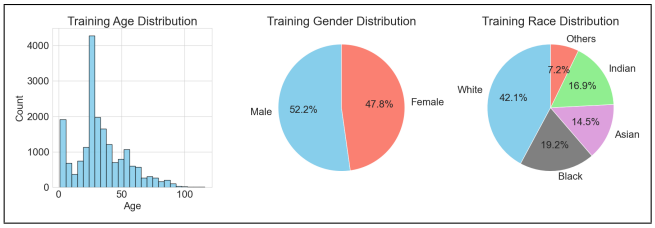

Figure 6. Training set visualizations of training set distributions for age, gender, and race.

5. Experiments and Results

In order to compare the performance of these methods to the baseline model, we used a fairness metric inspired by

	Age	Gender	Race
count	19041	19041	19041
mean	33.346	0.477	1.274
std	19.926	0.499	1.346
min	1.0	0.0	0.0
25%	23.0	0.0	0.0
50%	29.0	0.0	1.0
75%	45.0	1.0	2.0
max	116.0	1.0	4.0

Table 1. Training dataset statistics for age, gender, and race labels.

Equal Opportunity Difference (EOD) as formally described by Feldman and Peake [13]. The authors define this metric for the case of a binary classifier with two protected classes, Male and Female (Equation 10). EOD measures the difference in the true positive rates between samples from the two protected classes. Ideally, the EOD for a perfectly fair classifier should be 0.

$$EOD = P(\hat{Y} = 1 | A = \text{male}, Y = 1) - P(\hat{Y} = 1 | A = \text{female}, Y = 1) \quad (10)$$

However, in our case of evaluating fairness across different races, our protected classes are the five racial categories labeled in the UTKFace dataset. Further, while our gender classification task is also binary, we need our modified fairness metric to apply to the age regression task as well. Given these requirements and the original characteristics of the EOD metric, we chose to use the standard deviation of performance over the five protected classes, which we will refer to as Protected Standard Deviation (PSD), as our fairness metric (Equation 11). For the gender classification task, "performance" refers to the ratio of correctly predicted samples to total samples of a given protected class (Equation 12). For the age regression task, "performance" refers to the Mean Average Error, equivalent to the previously mentioned L1 loss for a protected class (Equation 13). PSD is a reasonable extension of EOD, measuring the average difference in performance between protected classes with an ideal minimum value of 0.

$$PSD = \sqrt{\frac{\sum_{r=0}^{4}(P_r - \overline{P})^2}{5}} \quad (11)$$

$$P_r = \frac{\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = y_i, \text{race}_i = r)}{\sum_{i=1}^{N} \mathbf{1}(\text{race}_i = r)} \quad (12)$$

$$P_r = \frac{\sum_{i=1}^{N} |\hat{y}_i - y_i| \cdot \mathbf{1}(\text{race}_i = r)}{\sum_{i=1}^{N} \mathbf{1}(\text{race}_i = r)} \quad (13)$$

An effective bias reduction method should produce a lower Protected Standard Deviation than the baseline model, while maintaining a comparable overall accuracy or MAE across all classes. We evaluated this balance using the product of the appropriate performance metric (MAE or the inverse of classifier accuracy) and the PSD. We will refer to this metric as the Performance-Variance product (PV). During the training process, we chose hyperparameters based on a model's validation PV, since it is a good indicator of a balance of performance and fairness.

5.1. Baseline Results

Both the baseline gender and age head were trained using the Adam optimizer [19] with a learning rate of 0.001 and a batch size of 64 images. We used a dropout rate of 0.4 for all dropout layers because it led to the best performance on the validation set. The gender head was trained for 23 epochs and the age head was trained for 31. The baseline model's performance on the test set is outlined in Table 2, organized by UTKFace race category, and includes the previously described PSD and PV metrics.

Race	Gender Accuracy	Age MAE (years)
White	0.909	7.894
Black	0.886	7.239
Asian	0.880	5.189
Indian	0.892	5.884
Others	0.872	5.406
Overall	0.896	6.906
PSD	0.012	1.061
PV	0.014	7.325

Table 2. Baseline Model Results

The gender classifier performs rather evenly across the five protected categories, while the performance of the age regressor varies more significantly between the class with the highest MAE, White, and the class with the lowest MAE, Others. Though less severe in the case of the gender classifier, the baseline model is clearly biased. Figure 7 compares the baseline model's performance (with MAE inverted so that a higher value corresponds to better performance) to the distribution of races in the training set. Unexpectedly, there is little correlation between the amount of training samples and performance. This suggests that source of the bias could be something more subtle than the availablility of training data.

5.2. Direct Variance Penalization

For the modified age regressor, we trained for 42 epochs with the same hyperparameters as the baseline model and we chose the parameter for this method, v (see Equation 5), to be 0.7 after finding that this value consistently produced a low PV product on the validation set while maintaining a

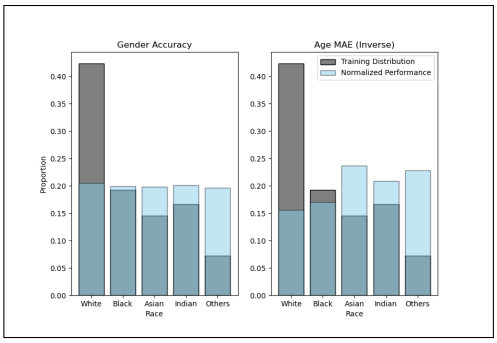

Figure 7. Baseline Performance vs. Training Distribution

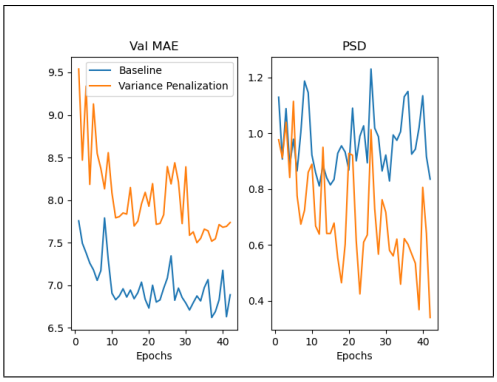

Figure 8. Variance Penalization Training Comparison

competitive MAE.

The modified gender classifier was trained for 50 epochs with the same value for v. Table 3 shows the test set results for both heads trained with the direct variance penalization method.

	Gender Accuracy	Age MAE (years)
Overall	0.896	7.796
PSD	0.020	0.712
PV	0.022	5.553

Table 3. Direct Variance Penalization Results

The modified gender classifier has the same overall accuracy as the baseline model with noticeably worse fairness across the five protected classes, evident in higher PSD and PV values. This trend will continue through our other bias minimization methods, which we attribute to the fact that the performance of the baseline classifier is already relatively uniform and any more sophisticated training process will only hinder the model.

However, this method significantly improves the fairness of the age regressor. The direct variance penalization method reduces the variance in MAE across protected races from 1.061 to 0.712, an improvement of nearly 33%. At the same time, the overall MAE is only raised by less than a year, resulting in a greatly improved PV product. Visualizing and comparing the baseline and modified training processes (Figure 8) confirms that, at the cost of slightly worse overall performance, the variance penalization method successfully improves the model's fairness throughout the training process.

5.3. Loss Scaling

For the loss scaling method, we trained the gender head for 38 epochs and the age head for 44 epochs with the same standard hyperparameters as the previous models. This method involves an additional moderation hyperparameter m, for which we used a value of 3 for the gender classifier and 5 for the age regressor based on the validation PV metric. It is logical that the optimal m value is smaller for the gender head than for the age head, since the baseline gender model already performs well in terms of fairness, so a less drastic scaling should be required. The loss scaling method's results are summarized in Table 4.

	Gender Accuracy	Age MAE (years)
Overall	0.881	7.024
PSD	0.013	0.832
PV	0.015	5.844

Table 4. Loss Scaling Results

Once again, although the modified gender classifier showed minor improvements to fairness on the validation set during training, these results did not generalize to the test set, where the gender head performs slightly worse than the baseline. Despite this, the modified age regressor successfully achieves a lower PSD than the baseline model while having only marginally worse overall MAE. Plotting the MAE and PSD throughout the training process (Figure 9) reveals that, unlike the direct variance penalization method, fairness is not learned over time, the PSD instead reaches a lower value than the baseline in a few epochs, and then PV improves as overall MAE is improved.

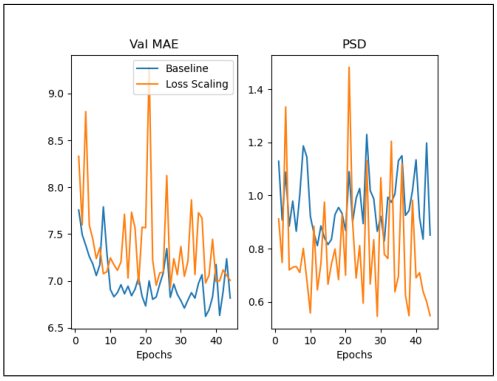

Figure 9. Loss Scaling Training Comparison

Qualitatively, the loss scaling training process seems to be noisier than that of the direct variance penalization

method. This is likely due to the re-scaling of the loss function to favor only one or two race categories.

5.4. Hierarchical Predictor

We trained the hierarchical model's internal race predictor for 41 epochs with the same standard optimization configuration as previous models. However, we used a higher dropout rate of 0.6 to combat a worse overfitting than we noticed with the other models. Table 5 summarizes the race predictor's test set accuracy among the five protected race classes. The model clearly struggles with the Others category, leading us to adjust the test-time behavior to automatically predict Other if the predicted race's confidence score is too low. The second row of the table shows the improved performance when using an "Others" threshold of 0.5.

	White	Black	Asian	Indian	Others
No Threshold	0.888	0.773	0.637	0.665	0.148
Threshold = 0.5	0.815	0.682	0.569	0.526	0.517

Table 5. Race Predictor Accuracy

Since the baseline age regressor performed the worst for the White class, we used the baseline model as the White class's individual age head and trained the other four class's heads so that their race's validation MAE was about the same as the White class's baseline validation MAE. To accomplish this, the remaining four class heads only needed to be trained for a single epoch with the standard optimization hyperparameters. The Black head was trained with a learning rate of 0.0005, the Asian head with 0.000025, the Indian head with 0.00025, and the Others head with 0.000025. Once all five race-specific age regressors were trained, we adjusted the race predictor's threshold parameter so that the full hierarchical model achieved the best validation PV product. We found that a threshold of 0.7 was optimal.

The individual gender classifiers were trained using a similar approach. The baseline gender accuracy was worst for the Others class, so the baseline was used as that class's individual head and the other heads were trained to match the validation accuracy. The White head was trained for one epoch with a learning rate of 0.0001, the Black head for 2 epochs with learning rate 0.001, and the Asian and Indian heads were both trained for 5 epochs with learning rate 0.001. An "Others" threshold of 0.6 for the race predictor led to the lowest validation PV. The results for the hierarchical model for both objectives are summarized in Table 6.

Like the previous methods, the hierarchical model shows no improvement for the gender objective but significantly improves the fairness of the age regressor at the cost of a higher overall MAE. This method produces the highest overall MAE out of the three discussed so far, which is not surprising given that this method involves purposefully training worse predictors for races with good baseline performance. Figure 10 shows the three test images that suffered the largest increase in the predicted age error from baseline.

	Gender Accuracy	Age MAE (years)
Overall	0.881	8.245
PSD	0.015	0.738
PV	0.017	6.082

Table 6. Hierarchical Predictor Results

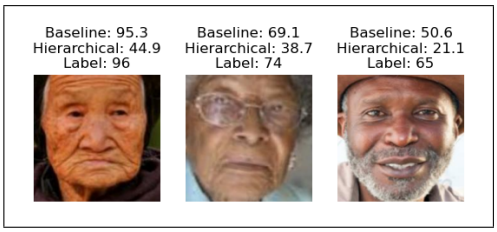

Figure 10. Largest Age Errors

5.5. Data Augmentation

We start data augmentation with calculating the target distribution, T_r, giving us how many more samples are needed for each race. Then we augment the dataset by randomly sampling and applying transformations to the existing training samples of that race. For each augmented image, a random transformation was selected from a specific set, including Horizontal Flip, Vertical Flip, 20 Degrees Rotation, 90 Degrees Rotation, Shear, Gaussian Blur, and Random Crop and Resize. These transformations were chosen to introduce variability while preserving the essential features of the facial images.

The focus for this section was the age prediction as we showed a much higher accuracy variation across race compared to that for gender classification. For the EfficientnetV2m base model, 256x256 FC gender classifier, 2048x2048 FC age regressor heads, after 31 epochs we measured a baseline training overall MAE of 6.7 years and by race MAE as 7.54, 6.41, 5.12, 6.65, 5.47, respectively.

Figure 11 shows the results for augmentation with moderation factor $\alpha = 1$ and $\nu = 1.5$. However we did not see much improved accuracy or PV product. This could be due to the large variations in the transformations applied and comparatively small augmented dataset.

Following the techniques used in Ref. [10], we decided to use a simpler augmentation with a random crop from any of the four corners followed by a horizontal flip. Figure 12 shows an example with the set of possible augmentations that we used and figure 13 shows these results with moderation factor $\alpha = 1$ and $\nu = 3$. We observed a positive impact where we can see that although we achieve a worse

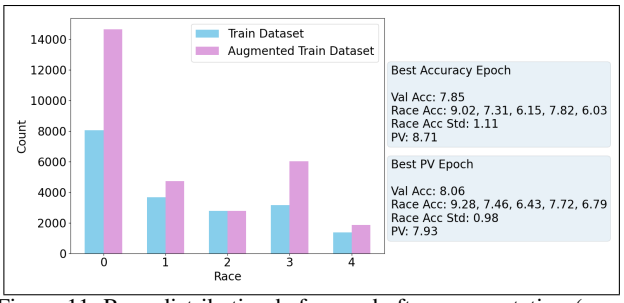

Figure 11. Race distribution before and after augmentation ($\alpha = 1$, $\nu = 1.5$) with random transformations. Age regression accuracy bias across races remains largely unchanged.

accuracy across different races, we are able to reduce the training bias. This is reflected through a reduced race accuracy standard deviation of 0.81 and consequently a smaller PV product value of 6.87.

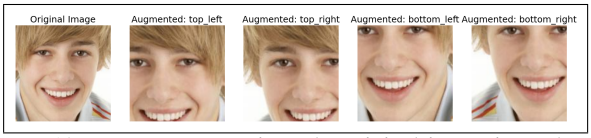

Figure 12. Image augmentation: the original image is randomly cropped from a corner, resized, and flipped horizontally.

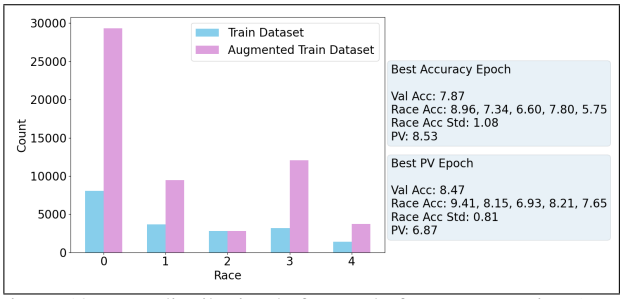

Figure 13. Race distribution before and after augmentation ($\alpha = 1$, $\nu = 3$) with random corner crop and horizontal flip. Accuracy decreased, but race accuracy variation and PV product for age regression were reduced.

6. Conclusions and Future Work

Table 7 summarizes the results of our four bias reduction methods compared to the baseline model.

In terms of the age regression objective, all four proposed methods successfully achieve a lower standard deviation of performance across the five protected race classes in the UTKFace dataset. Our methods were less successful on the gender classification objective. We attribute this to the fact that the baseline model performed quite fairly on the gender task, so equalization was not necessary and ended up only creating noisier models with poorer accuracy.

As we expected, our bias reduction methods improved on the baseline model's fairness at the cost of overall accuracy. In order to create more equitable performance, protected classes which the baseline model performs better on must be brought down to the level of the worst-performing class.

	Baseline	
	Gender Accuracy	Age MAE (years)
Overall	0.896	6.906
PSD	0.012	1.061
PV	0.014	7.325
	Variance Penalization	
Overall	0.896	7.796
PSD	0.020	0.712
PV	0.022	5.553
	Loss Scaling	
Overall	0.881	7.024
PSD	0.013	0.832
PV	0.015	5.844
	Hierarchical Predictor	
Overall	0.881	8.245
PSD	0.015	0.738
PV	0.017	6.082
	Data Augmentation	
Overall	N/A	8.471
PSD	N/A	0.810
PV	N/A	6.873

Table 7. Compiled Results

The Direct Variance Penalization emerges as the most effective of our four methods for the age regressor, boasting the lowest test PSD as well as only a modest increase to overall MAE, leading to the lowest PV product of 5.553. Based on these results, our fairest dual age/gender predictor would consist of the baseline gender head and an age head trained with the variance penalization method.

One of the largest limiting factors that we encountered over the course of our research was the size of the UTKFace dataset. With barely more than 20,000 labeled images, some of the baseline bias and weaknesses of our proposed methods could be attributed to the small dataset. In future work, we would like to take advantage of a larger dataset with age, gender and race labels if one becomes publicly available. We would also like to further develop the data augmentation method with more time and compute. Experimentation with this method was far slower than the other three because the augmentation prevented us from training on pre-saved features from the EfficientNetV2_M base CNN, which saved an enormous amount of compute.

7. Contributions and Acknowledgements

Jack Irish contributed to the development and implementation of the Variance Penalization, Loss Scaling and Hierarchical Predictor methods, as well as the design of training and testing tooling.

Dhruv Tandon contributed development and implementation of the data visualization techniques and the design and execution of the Data Augmentation method.

References

[1] D. Castelvecchi, "How facial recognition technology is shaping our world," 2020.

[2] S. Perkowitz, "The bias in the machine: Facial recognition technology and racial disparities," 2021.

[3] A. Najibi, "Racial discrimination in face recognition technology," 2020.

[4] V. Karimi and A. Tashk, "Age and gender estimation by using hybrid facial features," in *2012 20th Telecommunications Forum (TELFOR)*, pp. 1725–1728, 2012.

[5] S. Lapuschkin, A. Binder, K.-R. Muller, and W. Samek, "Understanding and comparing deep neural networks for age and gender classification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[6] I. Rafique, A. Hamid, S. Naseer, M. Asad, M. Awais, and T. Yasir, "Age and gender prediction using deep convolutional neural networks," in *2019 International Conference on Innovative Computing (ICIC)*, pp. 1–6, 2019.

[7] S. Hamdi and A. Moussaoui, "Comparative study between machine and deep learning methods for age, gender and ethnicity identification," in *2020 4th International Symposium on Informatics and its Applications (ISIA)*, pp. 1–6, 2020.

[8] A. Abdolrashidi, M. Minaei, E. Azimi, and S. Minaee, "Age and gender prediction from face images using attentional convolutional network," 2020.

[9] A. Garain, B. Ray, P. K. Singh, A. Ahmadian, N. Senu, and R. Sarkar, "Gra_net: A deep learning model for classification of age and gender from facial images," *IEEE Access*, vol. 9, pp. 85672–85689, 2021.

[10] P. Smith and C. Chen, "Transfer learning with deep cnns for gender recognition and age estimation," 2018.

[11] M. Du, F. Yang, N. Zou, and X. Hu, "Fairness in deep learning: A computational perspective," *IEEE Intelligent Systems*, vol. 36, no. 4, pp. 25–34, 2021.

[12] S. Wehrli, C. Hertweck, M. Amirian, S. Glüge, and T. Stadelmann, "Bias, awareness, and ignorance in deep-learning-based face recognition," *AI and Ethics*, vol. 2, pp. 509–522, Aug 2022.

[13] T. Feldman and A. Peake, "End-to-end bias mitigation: Removing gender bias in deep learning," 2021.

[14] M. Tan and Q. V. Le, "Efficientnetv2: Smaller models and faster training," 2021.

[15] A. Paszke, "Pytorch: An imperative style, high-performance deep learning library," pp. 8026–8037, 2019.

[16] Y. R. S. Marcel, "Torchvision the machine-vision package of torch," in *MM '10: Proceedings of the 18th ACM international conference on Multimedia*, pp. 1485–1488, Association for Computing Machinery, 2010.

[17] J. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[18] C. Harris, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.