Improving Facial Recognition Performance of Race Classification Using Bayesian Dropout on FairFace Model

Peter Tadrous

Saint John's University; Queens, New York, United States
peter.tadrous16@stjohns.edu

Abstract

Recent facial recognition techniques have tried to tackle the issue of imbalanced training data. Insufficient variability in training data has been shown to explain most of the reduction in performance when ResNet34 is presented with images balanced by race, and I will attempt to pinpoint exactly where any other biases may lie. There is significant performance improvement in models that are tasked with recognizing faces that are included in their training data, which is why it's crucial to include all races in the training data. However, many widely used open source datasets are not very diverse, and as Dulhanty et al. notes, bias in training data begets bias in model performance. Algorithmic auditing of commercial face analysis applications has uncovered disparate performance for intersectional groups across several tasks. Poor performance for darker skinned females by commercial face analysis APIs has been reported, along with lower accuracy in face identification by commercial systems with respect to "lower skin reflectance", i.e. darker skin, by researchers at the US Department of Homeland Security. Directly from LFW dataset's website: "Many groups are not well represented in LFW. For example, there [is] … a relatively small proportion of women. In addition, many ethnicities have very minor representation or none at all."

Introduction

In recent years, there has been much criticism of several commercial computer vision systems such as Microsoft and IBM for their discrepancies in accuracy across certain sub-demographics. These systems have found their way into many areas including security, medicine, education, and social sciences; but despite the vast amount of available data to train these systems, existing public face datasets such as LFW and Color FERET are strongly biased toward Caucasian faces, male faces, and adult faces, which significantly underrepresents minorities across all three of these sub-demographics.

Buolamwini et al. found that the commercial face gender classification systems all generally perform better on male faces and on light faces, a discrepancy which can often be attributed to biases in their training data.

It is well known that various unwanted biases in image datasets can easily occur due to biased selection and capture of these images, and Kärkkäinen et al. notes that the reason for the discrepancies in accurately classifying race is because most public large scale face datasets have been collected from popular online media – newspapers, Wikipedia, or web searches – and they assert that these platforms are more frequently used by or showing White people. Following the discoveries of Kärkkäinen et al., their team created a balanced dataset across sub-demographics, including race, age, and gender.

The primary focus thus-far has been on improving these systems before training, but I believe there is still room for improvement in the model itself. Gal et al. developed a method of implementing Dropout as a Bayesian Approximation to represent model uncertainty in Deep Learning. I expect this will yield much better performance as there is quite a bit of uncertainty in sub-demographic classification already, such as faces that are multi-racial or ambiguous.

Methodology

I began by recreating the ResNet34 model in tensorflow. I used a modified package to create the ResNet structure, ensuring there was no top layer. Following this, I imported the FairFace7 model in pytorch to copy over the weights and biases.

Here is where I began creating different models. The first model has no modifications to the original FairFace model, with the exception of the last layer having only 7 classes and using softmax activation. Since this applies to all three models, I will not make reference to this modification again.

The second augmented model has a traditional Dropout layer before the classifier, with a Dropout rate = 0.5. This augmented model is part of the experiment to see if there is significant change of performance not just between the control and the Dropout as a Bayesian Approximation, but also to show that changes in performance, if any, are in fact attributed to Bayesian Approximation or if it was just the addition of a Dropout layer that changed performance. I want to know if it is truly Bayesian Approximations that leads to better or worse performance.

Finally, the third model uses Dropout as a Bayesian Approximation directly before the classifier, to measure if there is any significant change in performance. The data used for training and testing is the FairFace cropped image dataset created using dlib's get_face_chip() to crop and align faces with padding = 0.25. The distribution of data across the different races is crucial to the efficacy of this experiment, and is shown in **Table 1**. I used pandas and shutil to read in the labels for the images to only focus on the Race label and move the images to their own subdirectories in the train and val directories for the image data generator.

To maintain the previous training of the original FairFace model, we will take advantage of the technique of transfer learning. Since this is a large model with nearly 100,000 images across training and validation, transfer learning will occur in 3 phases.

- Phase 1: Only the added classifying layers will be trainable.
- Phase 2: Only the ResNet34 layers without the added classifying layers will be trainable.
- Phase 3: All layers in the model will be trainable.

To ensure robust results after training, I used an image data generator to normalise the images, where I set the rotation_range = 10, set the width_shift_range = 0.2, height_shift_range = 0.2, and horizontal_flip = True. I utilized early stopping to minimize validation loss, and model checkpoint callbacks to maximize validation accuracy. Phase 1 of training used 30 epochs, Phase 2 used 20 epochs, and Phase 3 used 12 epochs.

Results

The results of all three models were quite poor compared to the original FairFace model. The FairFace model had an accuracy of 0.815 when classifying by race. The recreated FairFace model in tensorflow (model_1) had a **validation accuracy of 0.626** and a **validation loss of 1.705**. It had a training accuracy of 0.936 and a training loss of 0.174.

The model with a Dropout rate = 0.5 (model_2) had a **validation accuracy of 0.612** and a **validation loss of 1.408**. It had a training accuracy of 0.869 and a training loss of 0.355.

Lastly, the model with a Dropout as a Bayesian Approximation (model_3) had a **validation accuracy of 0.617** and a **validation loss of 1.697** respectively. It had a training accuracy of 0.944 and a training loss of 0.154.

Discussion

As we can see in **Figure 1**, in going from the second training Phase to the third training Phase, there is definitely an over fitting to the training data. I suspect there were too many epochs, in total, across all 3 training iterations. However, we did see in going from the first training Phase to the second training Phase that the accuracy for both training and validation improved, and we saw that the loss for both the training and the validation decreased. Additionally, we saw that the addition of the **Dropout as a Bayesian Approximation (model_3) did not have a marked improvement in accuracy when compared to the control model (model_1)**, across all three training Phases.

I attribute these findings to the architecture of the model used. ResNet architecture does not use any Dropout layers for regularization, and instead uses pooling layers. In this experiment, my models made use of three different regularization techniques: Dropout, early stopping, and data augmentation. Further research may be able to explore the effects of any one of these techniques individually or even additional techniques.

References

Kärkkäinen, K. and Joo, J. 2019. *FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age.*

Chou, Hsin-Rung, Jia-Hong Lee, Yi-Ming Chan, and Chu-Song Chen. 2018. *Data-Specific Adaptive Threshold For Face Recognition And Authentication*.

Deng, Jiankang, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. 2019. *Retinaface: Single-Stage Dense Face Localisation In The Wild*.

Deng, Jiankang, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2018. *Arcface: Additive Angular Margin Loss For Deep Face Recognition*.

Dulhanty, Chris, and Alexander Wong. 2020. *Investigating The Impact Of Inclusion In Face Recognition Training Data On Individual Face Identification*.

Buolamwini, Joy and Gebru, Timnit. 2018. *Gender shades: Intersectional accuracy disparities in commercial gender classification*.

Wu, Wenying and Protopapas, Pavlos and Yang, Zheng and Michalatos, Panagiotis. 2020. *Gender Classification and Bias Mitigation in Facial Images*.

Gal, Yarin, and Zoubin Ghahramani. 2015. *Dropout As A Bayesian Approximation: Representing Model Uncertainty In Deep Learning*.

Table 1.

Race	Training Samples	Validation Samples	Total
Black	12,233	1,556	**13,789**
East Asian	12,287	1,550	**13,837**
Indian	12,319	1,516	**13,835**
Latino/Hispanic	13,367	1,623	**14,990**
Middle Eastern	9,216	1,209	**10,425**
Southeast Asian	10,795	1,415	**12,210**
White	16,527	2,085	**18,612**
Total	**87,744**	**10,954**	**97,698**

Figure 1.1: Phase 1 Training for ResNet34 FairFace Model (model_1).

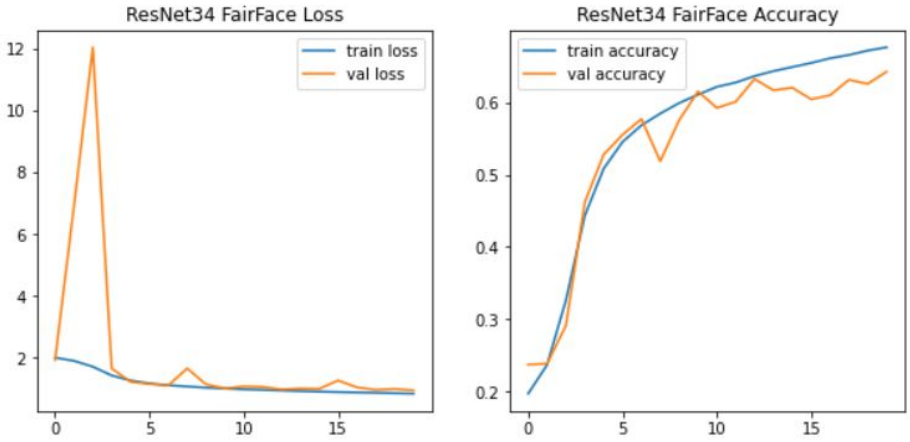

Figure 1.2: Phase 2 Training for ResNet34 FairFace Model (model_1).

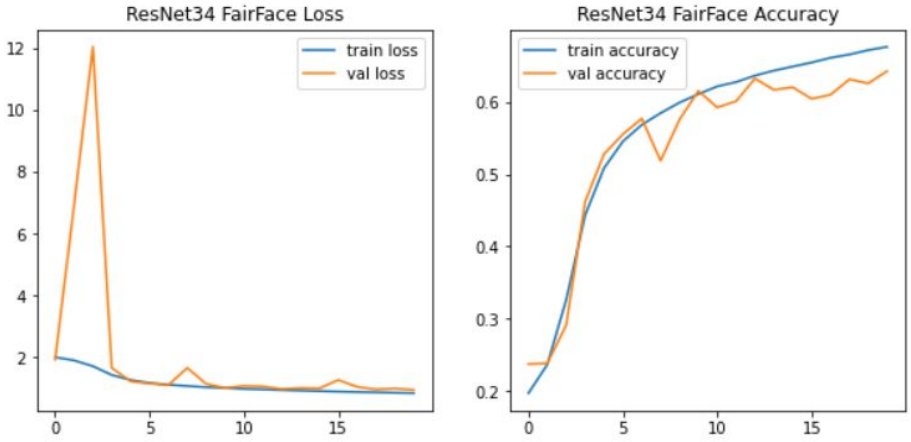

Figure 1.3: Phase 3 Training for ResNet34 FairFace Model (model_1).

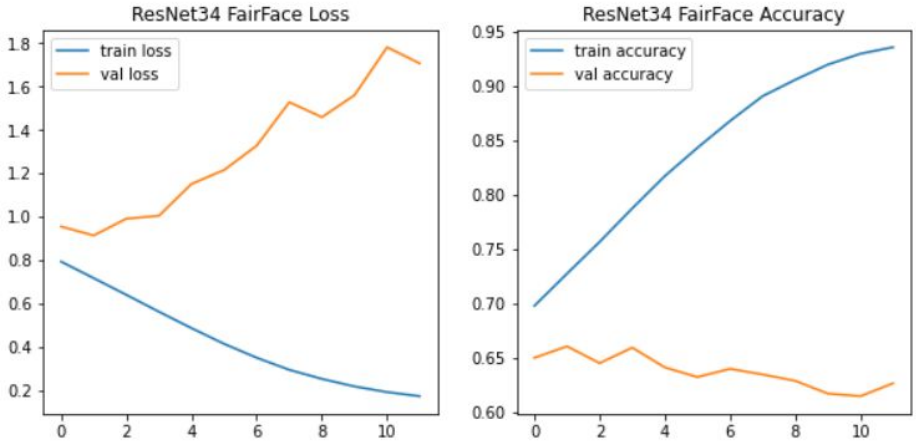

Figure 1.4: Phase 1 Training for ResNet34 FairFace Model with Dropout rate=0.5 (model_2).

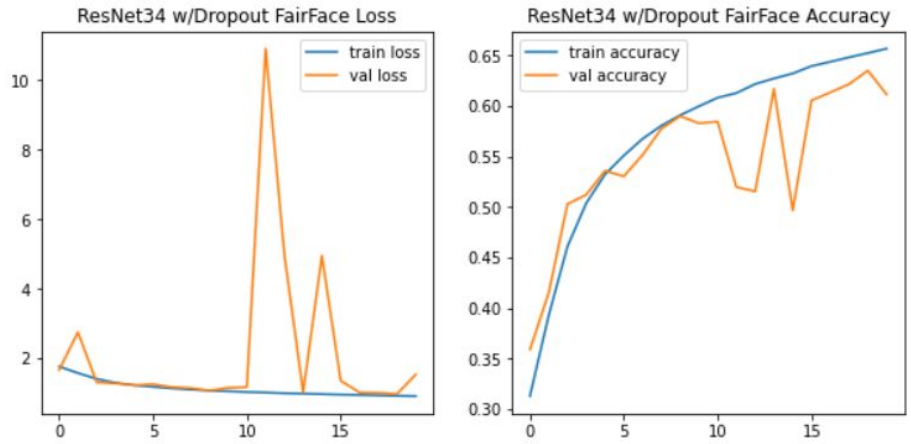

Figure 1.5: Phase 2 Training for ResNet34 FairFace Model with Dropout rate=0.5 (model_2).

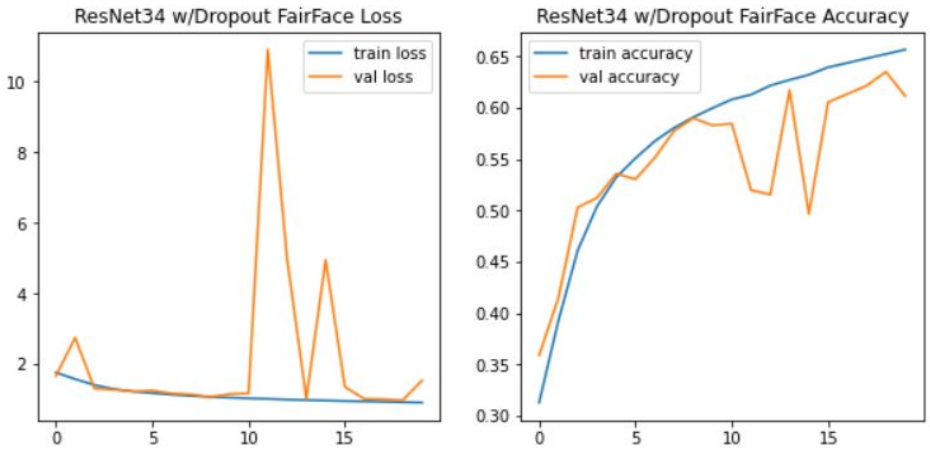

Figure 1.6: Phase 3 Training for ResNet34 FairFace Model with Dropout rate=0.5 (model_2).

Figure 1.7: Phase 1 Training for ResNet34 FairFace Model with Bayesian Dropout (model_3).

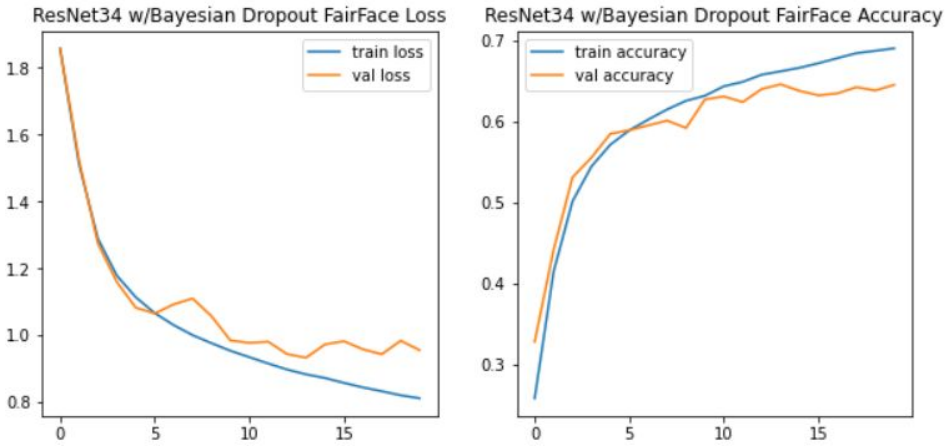

Figure 1.8: Phase 2 Training for ResNet34 FairFace Model with Bayesian Dropout (model_3).

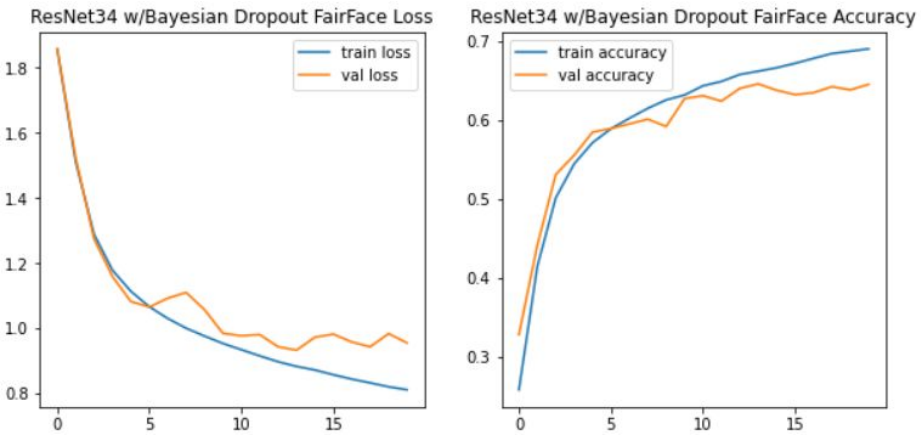

Figure 1.9: Phase 3 Training for ResNet34 FairFace Model with Bayesian Dropout (model_3).