FairNet: A Regularization Based Approach to Improve Fairness in Image Classification and Regression

Debolina Halder [1] Hailey Szadowski[1] Harshith Sesham [1] Noushin Quazi [2]

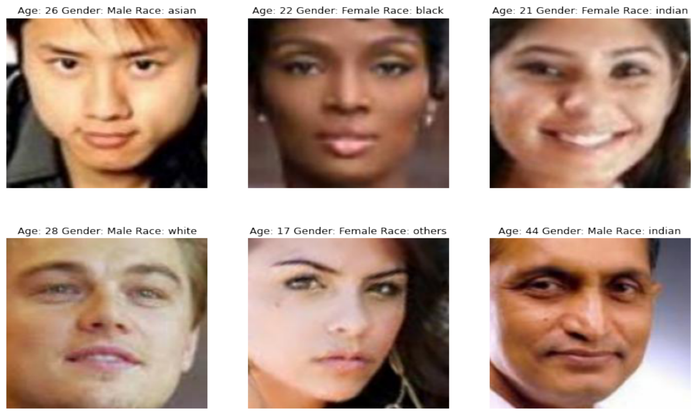
[1]Rice University [2]Baylor College of Medicine

Motivation

- Machine learning models are currently used to make critical decisions in society.
- Recent studies show that many of these models show biases against some protected groups like gender, race, age, or other sensitive attributes ([4]).
- Image classification is widely used in many critical sectors like automated inspection and quality control, object recognition in driverless cars, detection of cancer cells in pathology slides, face recognition in security, traffic monitoring and congestion detection, etc ([4]).
- Image regression is widely used in age estimation, property value evaluation, chemical analysis, healthcare, etc ([3]).
- Therefore, it is critically important to reduce the algorithmic bias in image classification and regression to ensure fairness ([1]).

Dataset

- The dataset used is the UTKFace dataset, a Kaggle dataset containing over 20,000 images of faces ([5]).
- We are predicting age and race.
- Gender is our protected attribute.



Bias

In literature, algorithmic fairness is often described with bias. The lower the bias the better the fairness. Let z be our protected group.

- Classification:

$$L_B(y, \hat{y}, z) = \left|p(y = \hat{y}|z') - p(y = \hat{y}|z)\right|$$

- Regression:

$$L_B(y, \hat{y}, z) = \left|\sqrt{E\left[(y - \hat{y})^2|z'\right]} - \sqrt{E\left[(y - \hat{y})^2|z\right]}\right|$$

Method

- Estimating age is a regression task and estimating race is a classification task.
- For both cases first we have trained a basic CNN model. Let the model be M and the loss be $L_M(y, \hat{y})$. Let us call this the base model.
- at each iteration, we are adding the bias associated with the outputs as a regularizer to the loss function L_M. Let the bias be $L_B(y, \hat{y}, z)$.
- We also use a regularizing hyper-parameter α to control the effect of the regularizer on the model.

Loss Function

$$L(y, \hat{y}, z, alpha) = L_M(y, \hat{y}) + \alpha * L_B(y, \hat{y}, z)$$

Model Architecture

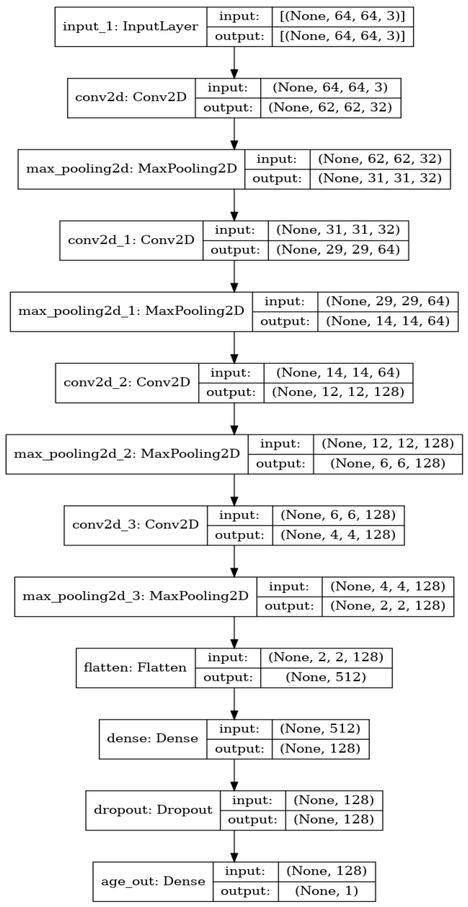


Figure 1. Architecture of Age Estimator

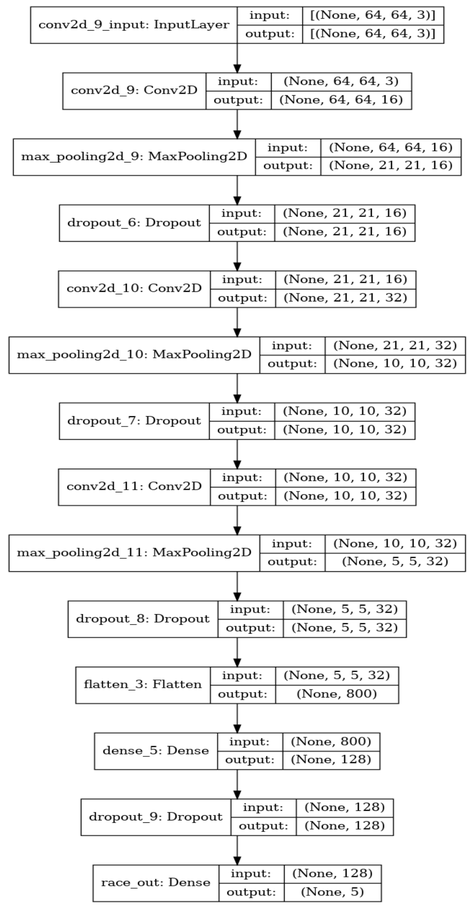


Figure 2. Architecture of Race Classifier

Results

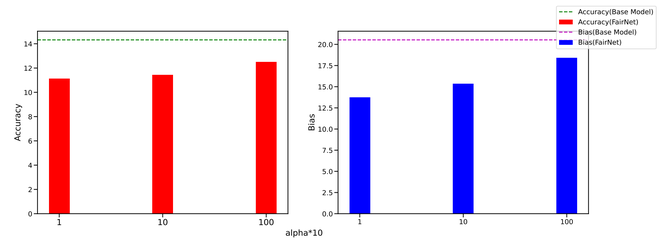


Figure 3. Bias and Accuracy of age estimation for different values of α

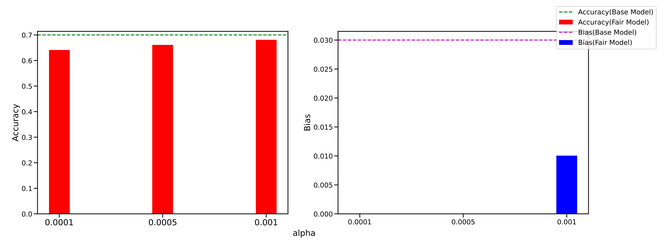


Figure 4. Bias and Accuracy of race classification for different values of α

Conclusion

- The regularizer improves fairness with some cost in accuracy.
- A larger value of α results in lower accuracy and higher fairness.
- For regression, the bias loss L_B depends on the range of output value. α should be chosen carefully.
- The age estimation model is more biased than the race classification model

Future Work

- Apply this method to CelebA dataset
- Apply this method to other classification and regression problems that do not involve images

References

[1] Xiaoxiao Li, Ziteng Cui, Yifan Wu, Lin Gu, and Tatsuya Harada. Estimating and improving fairness with adversarial learning. *arXiv preprint arXiv:2103.04243*, 2021.

[2] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[3] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

[4] Huan Tian, Tianqing Zhu, Wei Liu, and Wanlei Zhou. Image fairness in deep learning: problems, models, and challenges.

[5] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages