Comparison of the Combination of Bias Mitigation Methods

Bill Shao

September 2020

Contents

Abstract

Modern-day facial recognition networks utilize facial features to determine various characteristics of a person such as emotion, age, etc. Though many of these characteristics are intrinsically unrelated to other features such as the person's gender or race, studies have shown that networks will still correlate such features when training, which creates bias. Such design becomes problematic when considering the use of facial recognition in security and law enforcement, which can lead to tangible repercussions for minorities the system is biased against. Currently, many algorithms and methods exist to reduce such bias, from altering the dataset to be more representative to creating bias-aware or bias-blind algorithms. The three main methods considered in this study are (1) Strategic Sampling, (2) Adversarial Training, and (3) Domain-Independent Training. Though research exists comparing these three methods independently, this study seeks to explore the outcome when such approaches are combined and provide an in-depth analysis of their real-world impact. This study determines Adversarial Training combined with Strategic Sampling to be the best method in most cases, with slight variation considering the use of Domain-Independent Training for specific use-cases.

1 Introduction

In society, the push for the incorporation of machine learning into everyday tasks has greatly increased. Various machine learning tools are used to diagnose diseases, predict natural disasters, etc [18][12]. What is concerning, however, is the implementation of machine learning in fields subject to societal bias and pressure, such as facial surveillance, risk assessment for bank loans, and criminal bail. This is due to the fact that, although unintentional in most cases, existing ML models often inadvertently learn existing social bias, leading to discriminatory results [8][4][3]. In some application fields, such as voice recognition, bias poses a less serious issue. In others such as surveillance and loan assessment, however, bias can lead to adverse decisions. Biased bank loans or criminal bail charges can exacerbate ongoing economic inequality, and misidentifying someone in facial surveillance can lead to the pressing of false charges [2][13][19]. Looking at the adverse consequences of encoded bias in the mentioned fields, many are now seeking to produce various algorithms and datasets that can combat bias in the general field of machine learning fairness.

This study will primarily focus on the comparison between such algorithms, and also explore the possibility of combining them when possible to analyze the potential affect of it. Though applicable in many different cases, this study will specifically look at the implementation of such algorithms in Convolutional Neural Networks (CNN). The discussion of bias and racism in CNNs specifically is extremely high, and the trialing of facial surveillance technology is also significant, especially in countries such as China [4][8][15]. This work specifically targets and addresses this ongoing issue, but the general application case is generalizable. This is due to the fact that the three major methods **Strategic Sampling, Adversarial Training, and Independent Domain Training,** are all applicable to neural networks in general [20][6]. The latter parts of this study explores the subject of fairness metrics and parity in order to diagnose and assess algorithm usefulness. Usually, accuracy is an ineffective and misleading measure when diagnosing algorithmic fairness due to implicitly skewed datasets [7]. Should a dataset be heavily skewed in favor of Class A over Class B, for example, accuracy may only be a measure of how skewed the model is at predicting A, rather than its ability to discern A and B. This paper adds to current literature by bolstering existing knowledge and analysis of popular fairness algorithms, and also contributing the idea of combining multiple algorithmic approaches.

2 Related Work

Currently, much work has already been pursued and conducted in the field of Bias Mitigation regarding Convolutional Neural Networks. This study has mainly been motivated by societal implications regarding the use of CNNs for various tasks such as identification or prediction, where bias can have unintended consequences. Such biased systems have already been proven to exist and unfairly influence model decision making in CNNs as well as other systems[17][11]. Most modern-day responses either formulate dataset-augmentation approaches or algorithmic fairness approaches in order to reduce the occurrence of bias in such systems.

Datasets. Datasets such as FairFace, used for Facial Recognition, employ equal representation across genders and ethnic groups, meaning models are less incentivized to ignore minority groups or not be able to train for them [14]. This paper considers a variation of this approach known as strategic sampling (ss), which will instead weigh the losses of each result based on the category to which they correspond [20][6]. This will effectively simulate the effect of having an equally distributed dataset, but by using an algorithmic approach[20].

Algorithms. Aside from dataset modifications, algorithmic training methods such as Adversarial Training and Domain Independent training have also been proposed to counter bias [20]. Adversarial Training refers to a model learning a certain label X while also trying to prevent itself from learning the biased factor Y [20] [9]. This is done with 2 classifiers, one which predicts the label and another that predicts the biased label [5]. Domain Independent training divides the groups for label X by the biased factor Y, and instead tries to learn for each group individually EX: (Black Dog, Brown Dog, Black Cat, Brown Cat) vs. (Dog vs. Cat). Such approaches have been compared and considered independently in the past, but the effort to combine and analyze the result of using such methods in conjunction has not been explored. Many previous models are trained exclusively on one method, and also do not consider the use of a balanced dataset [20]. This paper seeks to utilize these prior implementations and also build on them to compare a greater variation of approaches.

Evaluation. Equally important as reducing bias is the ability to identify it, which cannot be tracked by metrics such as accuracy that are skewed towards the majority population the dataset. Many metrics have been proposed in order to identify bias, such as Equalized Odds, Statistical Parity, and Predictive Parity [7]. Each metric has a specific use case and can be helpful in drawing out specific information regarding the prediction patterns of the model. Since the chosen dataset is already heavily skewed, statistical parity and equalized odds, which aim for equality between answers is less preferred to metrics like Predictive Parity, which is more useful in representing the dataset unique predictions [7] [1].

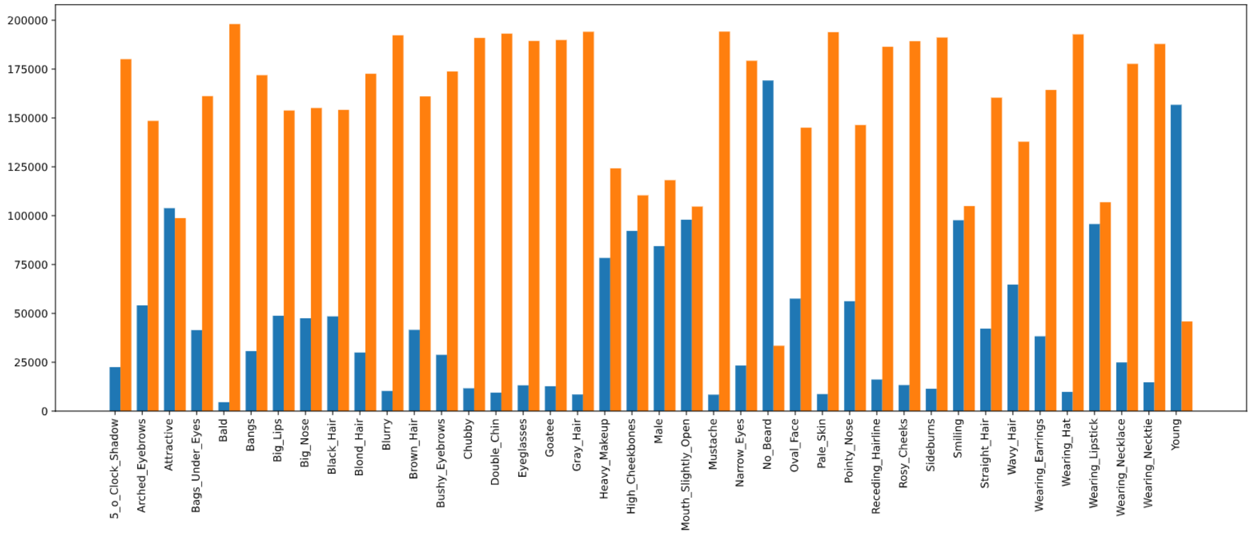

Figure 1: Bar Visualization of All Images with positive label (Blue) and with negative label (Orange).

3 Methods

3.1 Label Selection

CelebA provided a range of 20 total labels per image, 19 of which were considered as potential training metrics and the last, Gender, being used as the bias factor (Fig. 1) [16]. Out of the 19 considered labels, factors such as the overall amount of each label and relative proportion were considered to ensure both a large enough data sample to train and the ability to account for bias. Initially, labels such as “Bald, Gray Hair, Sideburns” were eliminated since there was too large a skew in the number of images without the label compared to with the label, meaning it did not serve as an adequate training set (Fig. 1). Other labels such as “Blurry, Attractive, Young” were also removed since they did not classify actual facial features or have objective definitions. Ultimately, this left the labels [Bags Under Eyes, Big Lips, Big Nose, Narrow Eye, Oval Face, Pale Skin, Pointy Nose, Smiling] as considerable options.

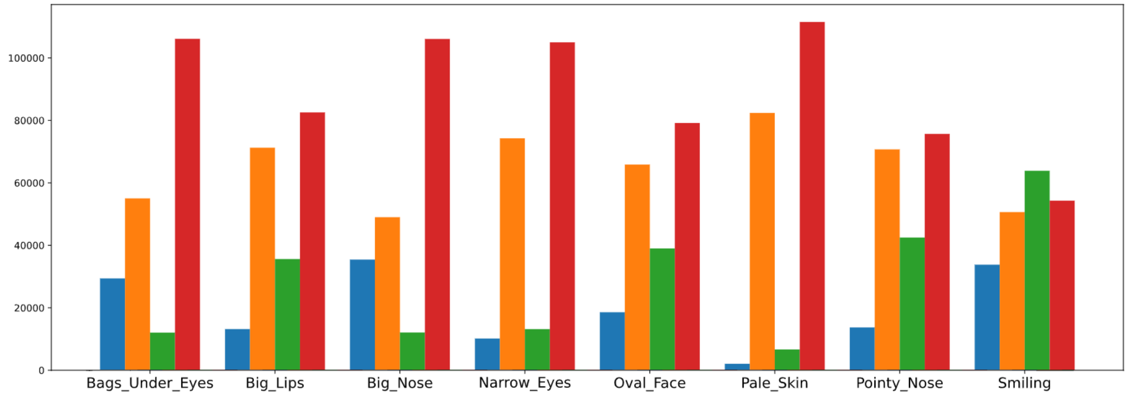

Figure 2: Bar Visualization of Narrowed Down Images with positive male label (Blue), negative male label (Orange), positive female label (Green), negative female label (Red).

From here, each label was divided again to consider the gender proportions alongside label value. The goal from this was the visualize which labels had differing gender ratios for positive and negative classifications, which should lead to a more biased baseline. For example, since the ratio of male_smiling : male_not_smiling is less than female_smiling : female_not_smiling, the model should more directly assume that females smile at greater rates than men, creating biased judgment [16] [17]. This consideration eliminated labels with extremely similar ratios such as pale_skin and narrow_eyes, which would make identifying bias more difficult. This narrowed down the potential labels to Bags_Under_Eyes, Big_Nose, Pointy_Nose, and Smiling. Though smiling initially seems the best choice because of its differing gender ratio, in order to also evaluate the affects of Strategic Sampling, labels with larger disparities between positive and negative categories were preferred [6]. Although a label such as Smiling may be useful for evaluating preset gender skew in data, it does little help interpret the affect of large positive : negative skews. Thus, big_nose was selected as the training label to identify bias because it had a large difference in gender ratio as well as a large positive to negative disparity.

3.2 Models Used

All code can be found at https://github.com/Billsha/CCBMM.

Adversarial Training. Adversarial Training utilizes a featurizer and two classifier components to both train for the big_nose label and also unlearn the gender label. The featurizer for Adversarial Training-training consists of a pre-trained Resnet-18 model without the classifier component, instead outputting a [512x1] tensor of features which is fed into two linear classifiers [10]. Both classifiers use the feature tensor to output a size [1] prediction for the big_nose label and gender label. The loss functions used for each classifier were Binary Cross-Entropy and outputs were normalized using sigmoid. In order to continually update and back-prop the model for both classifiers, a training_ratio constant was also set at a value of 3, meaning every third epoch would back-prop the network using the domain loss rather than the standard. On normal back-props, the model adds the standard class_loss and a confusion_loss which is calculated by the negative Sigmoid mean of the domain outputs. The confusion_loss acts to disincentivise learning for the domain_label, which would increase the confusion_loss and overall loss of the model.

Domain Independent. Domain Independent training has the same format as a standard CNN, using the pre-trained Resnet-18 model but modifying the FC-layer to be a linear layer [512x4], which represents the 4 possible outputs created with 2 possible variations for gender labels and 2 possible variations for big_nose labels [10]. Labels for the model were adjusted prior to training using a simple assignment method to have male_bignose = 0.0, female_bignose = 1.0, male_notbignose = 2.0, and female_notbignose = 3.0. Since this method varied in output type from the Adversarial Training model, Cross-Entropy Loss and Softmax functions were used in order to properly interpret the data.

Strategic Sampling. Strategic Sampling is a method of weighing the various subgroups in data to ensure equal importance when learning. This is done by passing an N-class weight matrix into the loss function prior to training, which will multiply the loss of each output by the corresponding prediction weight [20][6]. The weight matrix was created prior to training, and is meant to represent the distribution of the total dataset, rather than each individual batch. Combining this approach with the Domain Independent training method yielded a [1x4] weight matrix with each value corresponding to Male_Positive, Male_Negative, Female_Positive, and Female_Negative. The inverse of each percentile in [Fig 3.] was taken and normalized in order to create the matrix. Since the Adversarial Training training model has two classifiers rather than one, two weight matrices were used, both of size [1x2]. The class matrix held normalized inverse values of the big_nose distribution, and the domain matrix contained the same values for the gender label.

4 Results

Model	Accuracy	PPV (M)	PPV (F)	PPV Diff	NPV (M)	NPV (F)	NPV Diff
Benchmark	**85%**	0.693	0.343	0.350	0.633	**0.998**	0.365
Adversarial Training	83%	0.758	0.679	0.079	0.779	0.922	0.143
Adversarial Training (ss)	82%	**0.906**	**0.857**	**0.05**	0.836	0.944	**0.108**
Independent Domain Training	78%	0.732	0.519	0.214	0.821	0.941	0.129
Independent Domain Training (ss)	79%	0.788	0.606	0.182	**0.870**	0.997	0.127

Table 1: Calculated Results for Each Method. (ss): Strategic Sampling

$p(y|t) = \#$ of Images with predicted label y and true label t.

$$PPV = \frac{p(y=1|t=1)}{p(y=1|t=1) + p(y=1|t=0)}$$

$$NPV = \frac{p(y=0|t=0)}{p(y=0|t=0) + p(y=0|t=1)}$$

[7]

Model	Female Parity Difference	Male Parity Difference
Benchmark	0.665	0.06
Adversarial Training	0.243	0.02
Adversarial Training (ss)	0.087	0.07
Independent Domain Training	0.422	0.09
Independent Domain Training (ss)	0.391	0.08

Table 2: Difference in Gender Predicted Values to Assess Skew.

5 Discussion

Benchmark. First looking at accuracy scores for each model, the Benchmark predictably performed the best with a score of 85% in Table 1. This is expected specifically due to the demonstrated skew found in Figure 1, which shows the large gap between the big_nose groups and not big_nose groups. This means that the accuracy is representative of the Benchmark's ability to capitalize and learn the skewed patterns, which is also confirmed when considering it has the highest NPV for females. This most nearly means that it overwhelmingly predicts females as being without a big_nose, since the low number of $p(y = 0|t = 1)$ images in the dataset would essentially yield a NPV of 1/1 [7]. Conversely, this overwhelming skew can also be seen in the PPV for females, which was only 0.343, the lowest among the groups, due to its high $p(y = 1|t = 0)$ as a result of the heavily skewed predictions [7]. Similarly, the PPV for males is lower than the PPV for females, indicating similar trends but to a less extreme degree. This is most likely due to the fact that the with big_nose skew for males is less in Table 1. This is shown by the high gender parity difference in Table 2. Furthermore, the large values of PPV Diff and NPV Diff indicate a large inter-gender skew. This means, if implemented in society, the model would have the largest prediction disparity between men and women for both positive and negative predictions using the benchmark.

Adversarial. Expectantly, the Adversarial Training performed better than the Benchmark in terms of PPV Diff metrics and NPV Diff metrics, but suffered in accuracy as a result. When evaluating the NPV and PPV Differences, the significant improvement it has over the Benchmark also demonstrates the closing prediction gap between men and women. Especially notable is the PPV Diff of Adversarial Training with Stratified Sampling, which received a value of 0.05 in Table 1. This means that, in cases where the positive label is especially weighted and important, this model performs the best. For example, in the general context of surveillance or threat detection where a positive marker can lead to police deployment or escalation, this algorithm's low PPV difference means it is more unlikely to falsely detect males as threats or vice-versa. Though it shows significant improvement in this regard, it must also be noted that there are still trends of bias similar to ones in the Benchmark results, primarily in the comparison of the NPV (F) and PPV (F) results. In both the AT and AT with (ss) models, the NPV scored significantly higher, meaning the same logic that the model predicts more females as without big_nose applies in this example. Granted, this value is now much smaller, instead being 0.243 or 0.087 with (ss) compared to the 0.665 value from the benchmark in Table 2.

Adversarial w/ SS. In the measured statistics, the Adversarial Training with Stratified Sampling proves to be a more effective model, especially on the PPV Diff measurement. One interesting characteristic of these results, however, is the fact that PPV for Males is actually higher than NPV value for males, despite the fact that there are more males without big_nose in the dataset than with. This means the model actually categorizes males with a skew against the dataset pattern, which is potentially caused by the adversarial classifier aiming to reduce gender leaning [5].

Independent Domain Training. Interestingly, the IDT model produced the lowest accuracy and did not excel on many of the fairness metrics in consequence. The model itself only produced an accuracy of 75%, and scored low to mid tier on each fairness metric, being 4th and 3rd place on PPV Diff and NPV Diff respectively. This model also followed many of the existing trends, with a large skew towards NPV (F) and a slightly smaller one towards NPV (M) seen in Table 2. The Stratified Sampling metric also had a surprisingly small affect on IDT, only lowering the PPV Diff by 0.2 and hardly changing the NPV diff, though it did yield the best NPV (M) value (Table 1). This outcome is also relatively unexpected since weighting each subclass intuitively seems like the most favorable option in terms of fairer evaluation.

Comparison. Ultimately, Adversarial Training with Stratified Sampling seems to be the best overall model in terms of bias reduction, as it scores highly in PPV diff and NPV diff without a high discrepancy in the Gendered Parity Differences. This contradicts previous comparisons, though this difference could potentially be accounted for by the dataset label and model [20][14]. Some consideration can also be given to Independent Domain Training when considering negative label importance, though most likely only in unique cases. In this case, since Stratified Sampling was not significantly effective in both Adversarial Training and Independent Domain Training, it also cannot be concluded whether the method is beneficial in all cases or only specific models.

6 Conclusion

This paper compared the implementation of Adversarial Training and Independent Domain Training in combination with Stratified Sampling. Ultimately, Adversarial Training with Stratified Sampling was determined to be the best approach in terms of bias reduction. Looking forward, many questions still exist regarding these results and their implications. For example, the specific influence of Stratified Sampling is still not clear, as it was both successful and created little change in results. Also, though these bias reduction methods can be proven for the big_nose attribute, it is not yet clear if such results will translate into their implemented counterparts. The model itself is another unaccounted factor, and different networks may respond differently to both bias in general and the addition of reduction programs. The answer to these questions will hopefully push forward the goal for fairer machine learning.

References

[1] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H. Chi. Putting fairness principles into practice: Challenges, metrics, and improvements. *CoRR*, abs/1901.04562, 2019.

[2] Dan Ennis Cook and Tim. Bias from ai lending models raises questions of culpability, regulation, Aug 2019.

[3] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR*, abs/1808.00023, 2018.

[4] Kade Crockford. Aclu news amp; commentary, Jun 2020.

[5] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. *CoRR*, abs/1104.3913, 2011.

[6] Charles Elkan. The foundations of cost-sensitive learning. In *In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 973–978, 2001.

[7] Pratyush Garg, John Villasenor, and Virginia Foggo. Fairness metrics: A comparative analysis. *ArXiv*, abs/2001.07864, 2020.

[8] Karen Hao. A us government study confirms most face recognition systems are racist, Apr 2020.

[9] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[11] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. volume 108 of *Proceedings of Machine Learning Research*, pages 702–712, Online, 26–28 Aug 2020. PMLR.

[12] Naveen Joshi. How ai can and will predict disasters, Mar 2019.

[13] Aaron Klein. Reducing bias in ai-based financial services, Jul 2020.

[14] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age, 2019.

[15] Seungha Lee. Coming into focus: China's facial recognition regulations, Sep 2020.

[16] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[17] Daniel McDuff, Roger Cheng, and Ashish Kapoor. Identifying bias in ai using simulation, 2018.

[18] Paul Sajda. Machine learning for detection and diagnosis of disease. *Annual Review of Biomedical Engineering*, 8(1):537–565, 2006. PMID: 16834566.

[19] ACLU Staff. With ai and criminal justice, the devil is in the data, Mar 2019.

[20] Zeyu Wang, Klint Qinami, Yannis Karakozis, Kyle Genova, P. Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8916–8925, 2020.