Attribute Aware Filter-Drop for Bias-Invariant Classification

Shruti Nagpal[1], Maneet Singh[1], Richa Singh[2], and Mayank Vatsa[2]
[1]IIIT-Delhi, India; [2]IIT Jodhpur, India
{shrutin, maneets}@iiitd.ac.in, {richa, mvatsa}@iitj.ac.in

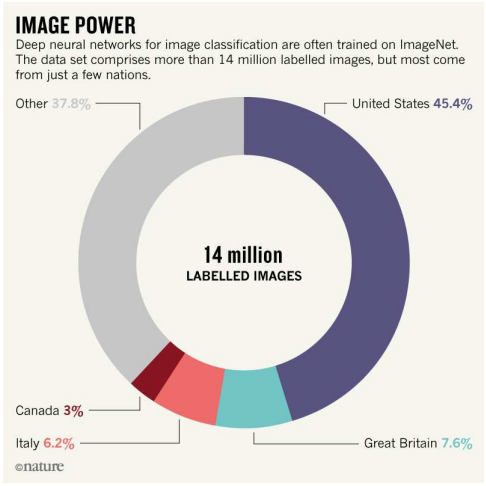
Abstract

The widespread applicability of deep learning based algorithms demands dedicated attention towards ensuring unbiased behavior. Biased feature learning (for or against a particular sub-group) might often result in unfair predictions. In order to address the above issue, this research proposes a novel Filter-Drop algorithm for learning unbiased representations. The proposed technique focuses on learning the features useful for predicting the biasing attribute (or the sensitive attribute), followed by their elimination while performing the primary classification task. To this effect, a multi-task network is trained, which prevents the features capturing the attribute variations from being used for the primary classification task. The efficacy of the proposed Filter-Drop technique is demonstrated on two facial analysis datasets: UTKFace dataset and FairFace dataset. The proposed technique achieves similar performance across different ethnicity groups while training with highly skewed training data as well.

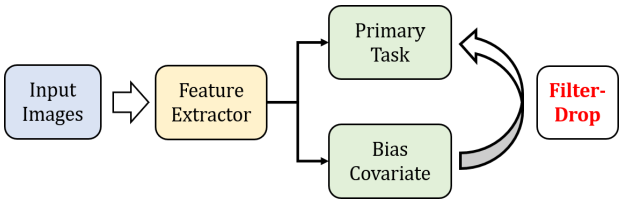
1. Introduction

Recently, the machine learning community has been marred with the challenge of bias and fairness in AI systems [4, 10, 14, 17, 19, 21]. Researchers have presented several case studies demonstrating bias in various applications and problems such as face recognition, attribute prediction, activity recognition, and automated caption generation. Today, machine learning systems are omnipresent in our lives, therefore it is essential to develop unbiased models, which do not present unfair outcomes. The predictions of these learned models should not be based on or biased against a particular sensitive attribute, thus resulting in bias-free outcomes. Unfair outcomes can have severe implications depending on the task of the machine learning models. For example, an automated recruitment tool should make hiring decisions based solely on the professional qualifications, without any inherent bias due to some other factor such as gender or age-group.

Existing studies have shown that the presence of bias can either be attributed to (i) the tainted examples which promote human bias against a particular sub-group, or (ii) skewed training samples which lead to imbalanced data with respect to a particular sub-group (Figure 1(a)) [3]. In computer vision, bias has usually been observed due to skewed training datasets. For instance, for a facial analysis model, the training samples might not be balanced with respect to an attribute such as gender or ethnicity.



(a) Distribution of images in the ImageNet dataset



(b) Proposed Filter-Drop Technique

Figure 1: (a) Location-wise distribution of the samples in the commonly used ImageNet dataset [7] [Source: https://www.nature.com/articles/d41586-018-05707-8]. (b) Diagrammatic overview of the proposed Filter-Drop technique. A multi-task network is learned to facilitate learning of bias-invariant features with respect to a given attribute.

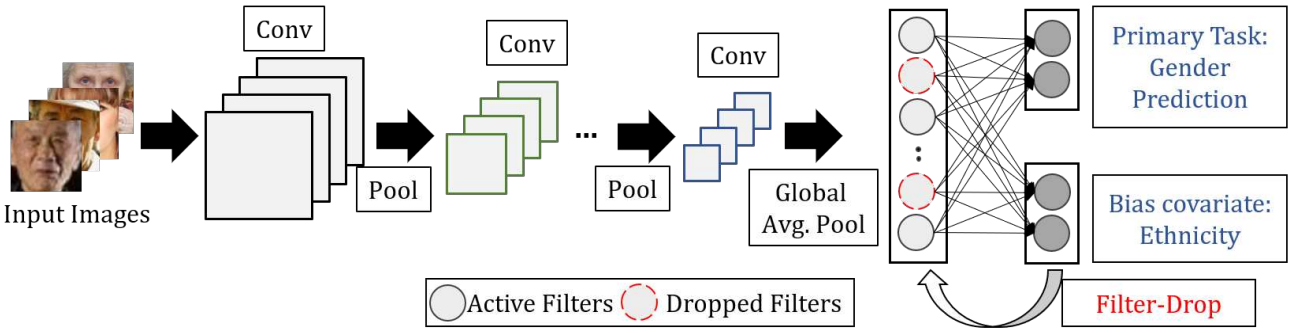


Figure 2: Diagrammatic representation of the proposed Filter-Drop algorithm for gender classification, under the sensitive attribute of ethnicity. A multi-task network is learned for gender and ethnicity prediction, such that the filters that are meaningful for ethnicity classification are dropped for gender prediction.

In this research, we propose a novel learning algorithm, which at the time of training *unlearns* the dependence of the model on sensitive attributes. These sensitive attributes are referred to as *bias co-variate*. The proposed algorithm, *Filter-Drop*, removes the convolutional filters responsible for encoding a given sensitive attribute (Figure 1(b)). For instance, consider the problem of gender prediction from face images with ethnicity as the sensitive attribute. During training, the data contains images belonging to different ethnicities, the proposed algorithms drops filters containing the ethnicity information (sensitive attribute), in order to make the predictions independent of it. The filters to be dropped are learned at the time of training via a multi-task network (Figure 1(b)) capable of performing the primary task (gender prediction) along with a secondary task of bias co-variate prediction (ethnicity classification). Filter-drop facilitates the removal of ethnicity features for gender prediction, thus promoting unbiased predictions. The efficacy of the proposed technique is demonstrated on two datasets: UTKFace [28] and FairFace [11] datasets. In order to simulate the real world scenarios, experimental evaluation is performed with skewed data as well as equal data (with respect to the sensitive attribute) to demonstrate the effectiveness of the proposed approach in eliminating the bias present in training samples.

1.1. Related Work

In the literature, researchers have proposed different techniques for addressing the problem of biased predictions in automated classification models, while also analyzing the biased predictions in different use-cases. Most of the techniques either focus on learning a bias invariant model, or attempt to debias previously trained models.

Researchers have attempted to define the concept of 'fairness', and have proposed techniques to incorporate the same in classification models [25]. Creager *et al.* [6] presented a technique to obtain flexibly fair features via a disentangled representation learning technique. Kim *et al.* [13] proposed a regularization algorithm for learning with biased data by optimizing over the mutual information between the feature embeddings and the bias. Alvi *et al.* [1] presented a domain adaptation based approach to debias neural networks at the time of feature learning by comparing classifiers trained on data containing spurious variations and a uniform distribution. Adversarial training paradigm [24] has also been utilized for learning bias-invariant models, while learning features independent of a given sensitive attribute [26]. Adversarial learning based techniques have also been presented for debiasing existing classification models [16, 22, 27]. Dwork *et al.* [8] proposed a technique which can be attached to existing black-box models for group-fair classification, wherein decoupled classifiers are learned. Amini *et al.* [2] proposed using a re-weighting based training algorithm for modifying the weights of an existing model in order to debias the model. This research focuses on the domain of learning bias-invariant features for facial analysis, with respect to a sensitive attribute. The proposed technique eliminates the features encoding information related to the sensitive attribute, in order to learn a bias-invariant model.

2. Proposed Attribute Aware Filter-Drop

In this research, we present an approach to perform bias-invariant and effective classification by learning to drop filters which promote dependency on an underlying attribute. Figure 2 presents an overview of the network architecture and the proposed Filter-Drop technique.

2.1. Filter-Drop

The proposed concept of filter-drop is similar to that of dropout [20]. Dropout has been used in literature for various applications [18, 23], and several variants of dropout have been proposed in the literature [12]. However, unlike dropout which is performed to improve the generalizability

of a network, filter-drop is performed with the specific aim of un-learning. A few filters are dropped or not used for training the subsequent layer, and therefore do not contribute to the final predictions made by the model. The filters to be dropped are learned during training, and the predictions are performed without the dropped filters. In this work, we perform filter-drop before the fully connected (FCN) layer of a network, after applying the global average pooling operator [15]. For an input x, a feature vector is obtained after the final convolution layer ($f(x)$), which is of dimension $d \times m \times m$, i.e., it consists of d filter activations, of dimension $m \times m$. A $d \times 1$ dimension vector is obtained upon applying the global average pooling operator to the feature maps. This is mathematically expressed as:

$$y = \phi(f(x)) \tag{1}$$

where, ϕ represents the process of global average pooling. Further, given the output of the global average pooling layer (y), we propose applying filter-drop to obtain y_{drop}, which can mathematically be written as:

$$y_{drop} = m * y \tag{2}$$

where m is a $d \times 1$ dimension binary vector, and $*$ refers to element-wise multiplication. If $m = 0$, the filter is dropped and the value is not used for prediction, whereas if $m = 1$, the filter is not dropped. The value of m is determined based on a pre-defined constraint which decides whether a specific feature will contribute towards the final prediction or not. The pre-defined constraint for m can either be defined as random or learned at the time of training to determine the active filters.

2.2. Attribute Aware Filter-Drop

As explained in the previous sub-section, the proposed filter-drop can be performed either by using a pre-defined value of m or by defining a constraint to learn m in order to drop the filters intelligently. In this research, *attribute aware filter-drop* is proposed, where the filters containing sensitive attribute-specific information are dropped. This is performed by adding additional constraints during training. In order to learn attribute aware filter-drop, a multi-task network is learned, and predictions are performed for an additional task of attribute prediction, also referred to as *bias co-variate prediction* or the *secondary task*. The loss function ($\mathcal{L}_{Proposed}$) for training a multi-task network via the attribute-aware filter-drop is written as follows:

$$\mathcal{L}_{Proposed} = \mathcal{L}_{Primary} + \mathcal{L}_{Attribute} \tag{3}$$

Top 'n' filters which contribute the most to the prediction of the bias co-variate are dropped in order to eliminate the underlying effect of the attribute being predicted. These top 'n' filters are chosen based on the weighted activations

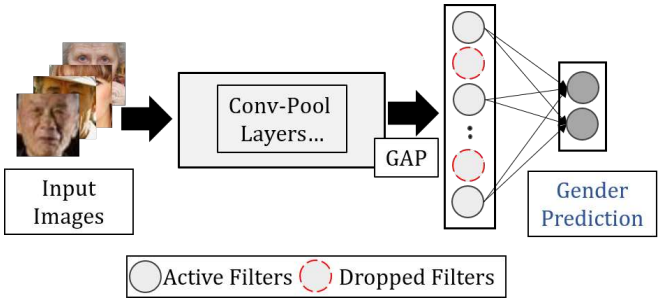


Figure 3: Using the attribute aware Filter-Drop model for gender classification during testing. The red filters are dropped and not utilized for the final prediction.

for the correct attribute class prediction (a higher weighted activation value refers to more contribution towards the final predicted label). Therefore, the predictions for the primary task are made using the $d - n$ filters, which contain limited correlation with the sensitive attribute. We can mathematically express this as:

$$m_i = \begin{cases} 0, & \text{if } i \in top\ n(y * W_{True-class}) \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

where, m_i corresponds to the i^{th} element of the vector m, $W_{True-class}$ is the final layer weight vector of the true attribute class for the corresponding input x. The above Equation is used to retrieve the top 'n' filters contributing the most towards the true class prediction. Thus, we attempt to eliminate the effect of the sensitive attribute for the primary task by limiting the information used for performing classification.

2.3. Bias-Invariant Classification

The proposed attribute aware filter-drop technique is presented in order to learn unbiased representations and thus perform bias-invariant classification. As shown in Figure 2, the model is trained as a multi-task network for the primary classification task and a secondary attribute prediction task. Here, the attribute corresponds to the sensitive attribute corresponding to which a network may be biased. Once the filters having the maximum weights for sensitive attribute prediction are identified, they are dropped to eliminate the effect of the sensitive attribute at the time of the primary classification task. The primary classification task is performed without the dropped filters, both at the time of training and testing. Since the filters are dropped from the penultimate layer, Filter-Drop ensures that the features encoding the sensitive attribute are not used for the primary task. The classification is performed with $d - n$ filters only where n represents the number of dropped filters that have learned the attribute specific information. It is important

Ethnicity: Asian

Ethnicity: Indian

Ethnicity: White

(a) UTKFace Dataset [28]

Ethnicity: Asian

Ethnicity: Indian

Ethnicity: White

(b) FairFace Dataset [11]

Figure 4: Sample images from the two datasets used for demonstrating variations across different ethnicity groups. The images are captured in unconstrained settings, often showcasing variations along the pose, resolution, lighting, and occlusions.

to note that during the test time, the network is a uni-task network with the objective of performing the primary classification task only (as shown in Figure 3).

3. Experiments and Implementation Details

The performance of the proposed attribute aware Filter-Drop technique has been evaluated for a facial analysis task, specifically, gender prediction. Given an input face image, a gender prediction model classifies the input either as *male* or *female*. Experiments have been performed on two datasets: (a) UTKFace dataset [28] and (b) FairFace dataset [11]. The UTKFace dataset contains over 20,000 face images from five different ethnicities with large variations across the pose, illumination, resolution, expression etc. Similarly, the FairFace dataset contains face images corresponding to seven ethnicities, collected from the Internet demonstrating a wide range of variations. For experiments, data corresponding to the Asian, Indian, and White ethnicity has been used from both the datasets. Figure 4 presents sample face images from the test set of the two datasets.

Two setups have been followed for both the datasets: (i) containing data belonging to the Indian and White ethnicity, and (ii) containing data belonging to the White and Asian ethnicity. In both cases, training has been performed by using equal data from both the ethnicities, and by using training data skewed towards a particular ethnicity. The above two protocols ensure that the proposed attribute aware Filter-Drop technique is evaluated on scenarios resembling the real world having the availability of limited imbalanced training data. For each setup, for the UTKFace dataset, a total of 4000 images have been used for training (containing equal number of male/female samples), while 13,000 images have been used for the FairFace dataset. The test set from the UTKFace dataset contains a total of 2,300 images and the FairFace test set contains 6,000 images, while ensuring equal ethnicity-wise and class-wise distribution for both the datasets.

3.1. Implementation Details

Experiments have been performed using the ResNet-50 architecture [9]. As demonstrated in Figure 2, the features from the final pooling layer are used for performing two tasks: (i) gender prediction (primary task) and (ii) ethnicity classification (secondary task). A single dense layer is attached to the final feature for the two tasks, respectively. Output of the top 100 filters are dropped during the gender classification for eliminating the component of ethnicity. The model is trained using the proposed attribute aware Filter-Drop technique for 50 epochs, using the Stochastic Gradient Descent optimizer with an initial learning rate of 0.01. After the initial 10 epochs, the dropped filters are estimated, which are updated after each consecutive epoch. The filters obtained at the final epoch are used during test time as well. The network weights are initialized with those learned on the VGG-Face2 dataset [5]. The proposed Filter-Drop has been implemented in the PyTorch framework.

4. Results and Analysis

Tables 1-3 present the results and analysis of the experiments performed using the proposed attribute aware Filter-Drop technique on the UTKFace and FairFace datasets. For all the experiments, the effectiveness of the proposed technique can be observed due to the lower accuracy variation observed between the test set of different ethnicities. Comparison has been performed with the native Softmax loss, termed as 'Traditional', where the ethnicity prediction branch is removed, and the model is trained for the single task of gender classification only.

Analysis Using Equal Training Data: Table 1 presents the performance of the algorithms when using equal training

Table 1: Performance of the proposed Filter-Drop technique for gender classification on two datasets when using equal training data from both the ethnicities. Two setups have been followed, where in the first, $\mathbf{E_A}$ refers to the Indian ethnicity, while $\mathbf{E_B}$ refers to the White ethnicity. In setup-2, $\mathbf{E_A}$ refers to the White ethnicity, while $\mathbf{E_B}$ refers to the Asian ethnicity. The proposed Filter-Drop technique demonstrates lower disparity between the performance of different ethnicities.

Dataset	Algorithm	$\mathbf{E_A}$	$\mathbf{E_B}$	Average
Setup-1				
UTKFace	Traditional	90.69	93.73	92.21
	Proposed	94.52	94.95	94.73
FairFace	Traditional	92.80	94.06	93.43
	Proposed	93.93	94.06	94.0
Setup-2				
UTKFace	Traditional	94.17	93.21	93.69
	Proposed	94.86	94.60	94.73
FairFace	Traditional	94.40	92.00	93.20
	Proposed	94.53	93.06	93.79

data from both the ethnicities. Performance is reported on the test sets of the UTKFace dataset and the FairFace dataset. Since the model has access to equal training data, the performance on the test set is expected to be near similar for different ethnicities. Across different setups, it is observed that the Softmax loss (traditional model) demonstrates relatively higher accuracy variation between the face images of E_A and E_B, while showing a variation of almost 3% between the performance obtained on both the ethnicities. On the other hand, the proposed attribute aware Filter-Drop technique demonstrates lesser disparity between the performance on the two ethnicities.

Analysis Using Skewed Training Data: Table 2 presents the ethnicity-wise performance for two setups on the UTKFace and FairFace datasets for gender prediction. Similar to the previous results, the Filter-Drop technique achieves a lower accuracy variation between the accuracy obtained on the two ethnicities, thus promoting bias-invariant model learning. In some cases, an accuracy variation of less than 1% (UTKFace dataset) is also observed, thus motivating the usage of the attribute aware Filer-Drop technique for learning attribute (ethnicity) invariant models.

Effect of Number of Filters: Table 3 presents the performance of the proposed technique obtained by varying the number of filters to be dropped. The performance is analyzed on the UTKFace dataset having skewed training data for Setup-1 (Table 2). The ResNet-50 architecture has 2048 filters in the last convolutional layer, and removing all the

Table 2: Gender prediction performance using skewed training data, where only 10% of $\mathbf{E_B}$'s data is used during training. Setup-1 utilizes images from the Indian ($\mathbf{E_A}$) and White ethnicity ($\mathbf{E_B}$), while setup-2 utilizes data from the White ($\mathbf{E_A}$) and Asian ($\mathbf{E_B}$) ethnicity. The proposed technique demonstrates improved performance and less variation across different ethnicities.

Dataset	Algorithm	$\mathbf{E_A}$	$\mathbf{E_B}$	Average
Setup-1				
UTKFace	Traditional	91.30	93.39	92.34
	Proposed	94.60	94.60	94.60
FairFace	Traditional	91.73	94.40	93.06
	Proposed	93.53	94.26	93.89
Setup-2				
UTKFace	Traditional	94.17	91.91	93.04
	Proposed	94.62	93.82	94.22
FairFace	Traditional	94.66	90.53	92.59
	Proposed	94.66	91.60	93.13

Table 3: Gender prediction accuracy (%) on two ethnicities with varying number of dropped filters using skewed training data for Setup-1.

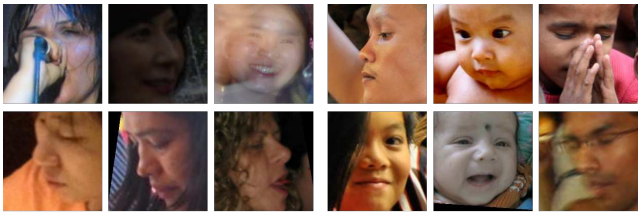
No. of Filters	$\mathbf{E_A}$	$\mathbf{E_B}$
50	94.69	93.91
100	94.60	94.60
250	94.52	94.69
500	94.68	93.73
2048	50.00	50.00

Table 4: Confusion matrix for gender prediction on the UTKFace dataset using skewed training data with Setup-1 (Indian and White ethnicities). The attribute aware Filter-Drop technique achieves similar performance across the two classes.

		True Label	
		Female	**Male**
Predicted	**Female**	1099	73
	Male	51	1077

filters results in random accuracy for the two class problem (50.00%). Lower variation is observed between the performance on the two ethnicities when removing 100 or 250 filters, as compared to removing 50 or 500 filters. It is our understanding that while dropping 50 filters does not result in the complete elimination of ethnicity information from the model, dropping 500 filters results in the loss of important discriminative information (useful for gender prediction).

Table 4 presents the confusion matrix for gender predic-



(a) Female (b) Male

Figure 5: Sample images from the FairFace dataset, misclassified by the proposed Filter-drop technique for gender classification. Large variations due to different covariates of resolution, pose, lighting, and occlusion render the problem further challenging.

tion on UTKFace dataset, when trained with skewed training data on Setup-1. The Filter-Drop technique demonstrates good classification performance across both the classes, where it achieves 95.56% for the female class and 93.65% for the male class. Further, Figure 5 presents sample images of the FairFace dataset which were incorrectly classified by the proposed Filter-drop technique. Most of the images suffer from large pose variations, resulting in limited captured face region. Certain images also demonstrate large variations due to the resolution or lighting of the image, thus making the problem further challenging.

5. Conclusion

Deep learning based facial analysis models have been shown to exhibit biased behavior, often resulting in incorrect predictions. Since such models are required to be used in social settings, with access to people from different subgroups, it is imperative to develop techniques which promote bias-invariant predictions. To this effect, this research proposes a novel attribute aware Filter-Drop technique for learning features invariant to a given attribute. Filter-Drop utilizes a multi-task network comprising of the primary task and a secondary task for bias co-variate prediction. The primary task refers to the main objective of the network such as facial analysis or object classification, while the secondary task corresponds to the classification of the biasing attribute. For example, for a gender prediction model, the primary task is to predict the gender of the given face image, whereas the secondary task is to predict the biasing covariate, i.e. ethnicity. The proposed technique extracts the top filters containing discriminative information with respect to the secondary task, and focuses on eliminating its features for the primary task. Elimination of the top filters results in the removal of the biasing factor (ethnicity) in the primary task predictions (gender). The performance of the proposed Filter-Drop technique has been demonstrated on two datasets: (i) UTKFace and (ii) FairFace. Over different experimental setups and varying training data distributions, the proposed technique demonstrates improved performance as compared to the existing algorithm. While current experiments utilize binary primary and secondary tasks, the proposed Filter-Drop technique can also be extended for multi-class problems.

6. Acknowledgement

S. Nagpal is partially supported by the TCS PhD Fellowship. M. Vatsa is partially supported through the Swarnajayanti Fellowship by the Government of India. R. Singh is partially supported through the Facebook's Ethics in AI Award.

References

[1] Mohsan S. Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In Laura Leal-Taixé and Stefan Roth, editors, *European Conference on Computer Vision Workshops*, pages 556–572, 2018.

[2] Alexander Amini, Ava P. Soleimany, Wilko Schwarting, Sangeeta N. Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *AAAI/ACM Conference on AI, Ethics, and Society*, page 289–295, 2019.

[3] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California Law Review*, 104:671, 2016.

[4] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 67–74, 2018.

[6] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning*, pages 1436–1445, 2019.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[8] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133, 2018.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[10] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 793–811, 2018.

[11] Kimmo Kärkkäinen and Jungseock Joo. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. *arXiv preprint arXiv:1908.04913*, 2019.

[12] Rohit Keshari, Richa Singh, and Mayank Vatsa. Guided dropout. In *AAAI Conference on Artificial Intelligence*, pages 4065–4072, 2019.

[13] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019.

[14] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[15] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

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

[17] Shruti Nagpal, Maneet Singh, Richa Singh, Mayank Vatsa, and Nalini Ratha. Deep learning for face recognition: Pride or prejudiced? *arXiv preprint arXiv:1904.01219*, 2019.

[18] Sungrae Park, JunKeon Park, Su-Jin Shin, and Il-Chul Moon. Adversarial dropout for supervised and semi-supervised learning. In *AAAI Conference on Artificial Intelligence*, 2018.

[19] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Conference on Fairness, Accountability, and Transparency*, page 469–481, 2020.

[20] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[21] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *European Conference on Computer Vision*, pages 498–512, 2018.

[22] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *IEEE International Conference on Computer Vision*, pages 5310–5319, 2019.

[23] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Conference on Computer Vision and Pattern Recognition*, pages 1249–1258, 2016.

[24] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In *International Conference on Neural Information Processing Systems*, page 585–596, 2017.

[25] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017.

[26] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

[27] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[28] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5810–5818, 2017.