FairGRAPE: Fairness-aware GRAdient Pruning mEthod for Face Attribute Classification

Xiaofeng Lin, Seungbae Kim, and Jungseock Joo ✉

University of California, Los Angeles, Los Angeles, CA 90024, USA
bernardo1998@g.ucla.edu, sbkim@cs.ucla.edu, jjoo@comm.ucla.edu

Abstract. Existing pruning techniques preserve deep neural networks' overall ability to make correct predictions but could also amplify hidden biases during the compression process. We propose a novel pruning method, Fairness-aware GRAdient Pruning mEthod (FairGRAPE), that minimizes the disproportionate impacts of pruning on different sub-groups. Our method calculates the per-group importance of each model weight and selects a subset of weights that maintain the relative between-group total importance in pruning. The proposed method then prunes network edges with small importance values and repeats the procedure by updating importance values. We demonstrate the effectiveness of our method on four different datasets, FairFace, UTKFace, CelebA, and ImageNet, for the tasks of face attribute classification where our method reduces the disparity in performance degradation by up to 90% compared to the state-of-the-art pruning algorithms. Our method is substantially more effective in a setting with a high pruning rate (99%). The code and dataset used in the experiments are available at https://github.com/Bernardo1998/FairGRAPE

1 Introduction

Deep neural networks (DNNs) are widely used in applications running on mobile or wearable devices where computational resources are limited [62]. A common strategy to improve the inference efficiency of deep models in such environments is model compression by pruning and removing insignificant nodes or connections between nodes, resulting in sparser networks than the original ones [4, 13, 14, 19, 34, 50]. These methods have been known to reduce the computational cost significantly with almost negligible loss in prediction accuracy [7].

Despite the prevalence of model compression, recent studies have also reported that compressed models may suffer from hidden biases, *i.e.* accuracy disparity, more severely than the original models [3, 25, 26]. The pruned models may be accurate overall or on some sub-groups (*e.g.* White males), while resulting more severe performance decrease from the original model on specific subgroups. This bias is particularly problematic for model pruning methods, which attempt to identify and remove insignificant parameters. The parameter-wise significance considered in such methods is estimated from in-the-wild datasets, which are typically unbalanced and biased [30, 51]. The societal impact of this

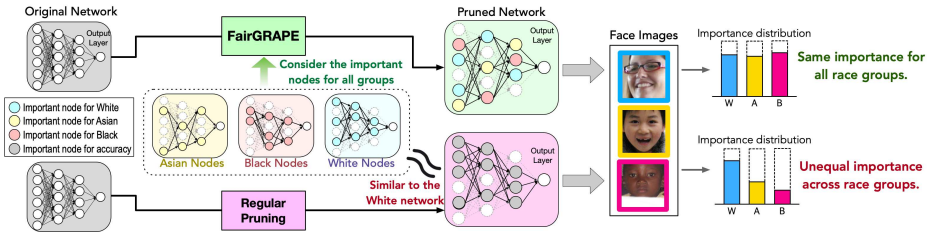

Fig. 1: Illustration of our proposed model compression method for face attribute classification. Since the compressed model pruned by a regular pruning method shows disparity results over different groups, our proposed pruning method aims to fairly treat all groups by preserving important nodes for sensitive attributes (*e.g.* race, gender) in networks.

bias is also huge because the compressed models are commonly used in consumer devices for daily use such as mobile phones and personal assistant devices.

To address this critical issue, we propose a novel model pruning method, Fairness-aware GRAdient Pruning mEthod – **FairGRAPE**. Our method aims at preserving per-group accuracy as well as overall accuracy in classification tasks. Figure 1 illustrates the fundamental idea of our proposed method. Existing pruning methods disregard demographic groupings and prune the nodes with the smallest weights to preserve the model's overall accuracy. However, some nodes may be critical only for a sub-population underrepresented in the dataset and consequently pruned, leading to a biased compressed model. In contrast, our method considers each node's importance to each sub-group separately so that it can retain important features for all groups.

Specifically, our method computes the group-wise importance of each parameter to get the distribution of the total importance of each group in a model. It then iteratively selects network edges that most closely maintain both the magnitude and share of importance for each group. By selecting such edges, our method equalizes the importance loss across groups, reducing performance disparity.

To evaluate the effectiveness of our method, we conduct extensive experiments on the face attribute classification tasks where demographic labels are readily available. We use four popular face datasets, FairFace [30], UTKFace [63], CelebA [38], and the person subtree in the ImageNet [56]. The experimental results show that FairGRAPE not only preserves the overall classification accuracy but also minimizes the performance gap between sub-groups after pruning, compared to other state-of-the-art pruning methods. We summarize our contributions as follows.

– We show that existing pruning methods disproportionately prune important features for different demographic groups, leading to a more considerable accuracy disparity in the compressed model than in the original model.

– We propose a novel, simple, and generally applicable pruning method that maintains the layer-wise distribution of group importance.
– We evaluate our method on four large-scale face datasets compared to four widely used pruning methods.

2 Related Work

2.1 Model Compression via Pruning

Compression of deep models involves various methods to reduce computation cost without significant loss in model performance. Major categories of compression techniques include Parameter pruning [13,20]; Parameter quantification [19]; Lower-rank factorization [47]; knowledge distillation [24]. In this paper, we focus on examining the first one: parameter pruning, which reduces the number of weights associated with nodes or edges in a network.

Prior research in pruning has focused on the following aspects: how to maintain certain structural elements of the original model [23, 35, 53], how to rank the importance of individual features [11, 34, 37, 40, 41, 50, 60], whether pruning should be done at once or across several steps [13, 59], and how many pruning and retraining iterations are required [4, 20].

2.2 Fairness in Computer Vision

Fairness has received much attention in the recent literature on computer vision and deep learning [5, 6, 16, 21, 29, 36, 39, 48, 49, 51, 52, 55, 57]. The most common goal in these works is to enhance fairness by reducing the *accuracy disparity* of a model between images from different demographic sub-groups. For example, a face attribute classifier may yield a disproportionately higher error rate on images of non-White or females [5, 30]. Another line of work has investigated biased or spurious associations in public image datasets and models between different dimensions of sensitive groups and non-protected attributes such as semantic descriptions, facial expressions, and age [1, 6, 28, 64, 65]. Our paper focuses on the former: the mitigation of accuracy disparity.

The cause of demographic bias can be demographically imbalanced datasets and the design choice of learning algorithms or network architectures [10, 32]. Prior works have found that a face dataset dominated by the White race produces a poor performance for other races, while a face dataset with balanced group distribution, from either real or synthesized data, can enhance fairness [15, 17, 18, 30, 52, 58]. Algorithmic bias can be mitigated through either explicit fairness constraints [31, 61], matching learned representations to a target distribution or group-wise characteristics [8, 44, 45], or adversarial mitigation and decoupling to disconnect representation and sensitive groups that attempts to decorrelate sensitive attributes and model outputs [1, 2, 12, 33, 43, 54]. Our method estimates the importance of each connection weight toward each sub-group and maintains between-group ratios in pruning.

2.3 Fairness in Model Compression

Only a few studies have been concerned with fairness in the compression of deep models. [25] reported that pruned models tend to forget specific subsets of data. It is examined in [26, 46] that pruning can impact demographic sub-groups disproportionately in face attribute classification and expression recognition. Another recent work [3] showed that knowledge distillation could reduce bias in pruned models. All these studies focus on measuring pruning-induced biases between output categories. To the best of our knowledge, our paper is the first to separate the pruning impact on output classes and sensitive groups and propose a pruning algorithm to mitigate biases in both dimensions.

3 Fairness-aware GRAdient Pruning mEthod

3.1 Problem Statement and Objective

Consider a neural network $\boldsymbol{f}$ parameterized by $\boldsymbol{\theta} \in \mathbb{R}^m$ and a dataset $D = \{(\boldsymbol{x_i}, \boldsymbol{y_i}, k_i)\}_{i=1}^n$, where $\boldsymbol{x_i} \in \mathcal{X}$ is an input vector, $\boldsymbol{y_i} \in \mathcal{Y}$ is a target output, and $k_i \in K$ is a sensitive attribute. The goal of network pruning is to find the following parameter set:

$$\boldsymbol{\theta'} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} L(D, \boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(\boldsymbol{\theta}, \boldsymbol{x_i}), \boldsymbol{y_i}), \text{ subject to } ||\boldsymbol{\theta}||_0 \leq c \cdot m \tag{1}$$

Here $\ell(\cdot)$ denotes a loss function, and $c \in (0, 1)$ is the desired sparsity level.

We further examine the network's performance on different subsets of D. Let $D_k = \{(\boldsymbol{x_i}, \boldsymbol{y_i}, k_i) | k_i = k\}$ denote the subset of instances from a sensitive group $k \in K$. Given a performance metric $A(D_k; \boldsymbol{\theta})$, the difference in performance on D_k between the full model $\boldsymbol{f}$ and a compressed model $\boldsymbol{f'}$ is:

$$\Delta A(D_k; \boldsymbol{\theta'}) = A(D_k; \boldsymbol{\theta'}) - A(D_k; \boldsymbol{\theta}) \tag{2}$$

The mean of all group-wise performance differences is:

$$\Delta \bar{A}(D; \boldsymbol{\theta'}) = \frac{1}{|K|} \sum_{k \in K} \Delta A(D_k; \boldsymbol{\theta'}) \tag{3}$$

Our goal is to minimize the variance of performance differences in a pruned model. This task can be formulated as finding the following $\boldsymbol{\theta^*}$:

$$\boldsymbol{\theta^*} = \underset{\boldsymbol{\theta'}}{\operatorname{argmin}}[Var(\Delta A(D_k; \boldsymbol{\theta'}))] = \underset{\boldsymbol{\theta'}}{\operatorname{argmin}}[\frac{1}{|K|} \sum_{k \in K} (\Delta A(D_k; \boldsymbol{\theta'}) - \Delta \bar{A}(D; \boldsymbol{\theta'}))^2] \tag{4}$$

Note that the actual task of the model determines the choices of the performance metric $A(\boldsymbol{\theta}; D_k)$. This paper focuses on classification tasks and thus uses accuracy, false positive rate (FPR), and false negative rate (FNR) as performance metrics. The output space $\mathcal{Y}$ and sensitive groups K can be either overlapping or disjoint, and this paper examines both cases.

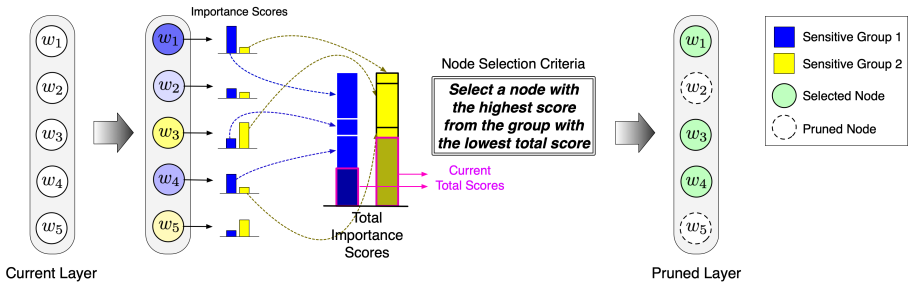

Fig. 2: Illustration of the proposed node selection method. FairGRAPE first computes the importance score of each individual weight for all groups layer-wise. Based on the total scores from the current layer, FairGRAPE selects a node with the highest score from the group with the greatest loss in importance score to minimize the variance of performance changes.

3.2 FairGRAPE: Fairness-aware Gradient Pruning Method

The common idea behind model pruning methods is estimating the importance of edges and pruning less important ones. While the existing methods focus on measuring the importance to the whole dataset, our method aims to preserve important weights for each sensitive group to mitigate biases.

To this end,we propose to compute the group-wise importance score of each weight with respect to each sensitive group, and then use a greedy algorithm to select weights based on the scores. At each step, the method compares the current ratio of importance scores with the target ratio (*i.e.* , the ratio in the model before the current pruning step). Then the group that has the largest difference will be selected, and the method adds one weight with the highest importance for the selected group to the selected network. Once the desired number of weights is selected, the remaining weights are pruned. FairGRAPE compresses all layers of the model with this node selection process, which is illustrated in Figure 2.

Group-wise Importance Let $w \in \boldsymbol{\theta'}$ denote a parameter in f' and $L(D_k, \boldsymbol{\theta'})$ denote the loss on sensitive group k. The gradient of $L(D_k, \boldsymbol{\theta'})$ with respect to w is $g_w = \frac{\partial L(D_k, \boldsymbol{\theta'})}{\partial w}$. Then the importance of w with respect to group k and the total model importance score for a group k are:

$$I_{k,w} = (L(D_k, \boldsymbol{\theta'}) - L(D_k, \boldsymbol{\theta'}|w = 0))^2 \qquad I_k = \sum_{w \in \boldsymbol{\theta'}} I_{k,w} \tag{5}$$

Computing the importance defined in equation 5 requires evaluating a different network for every parameter, which is often impractical. Alternatively, $I_{w,k}$ could be approximated by its first-order Taylor expansion, as explained in [41]:

$$I_{k,w} = (g_w w)^2 \tag{6}$$

Algorithm 1 FairGRAPE

$c \leftarrow$ desired sparsity
$r \leftarrow$ % of parameters to prune per iteration
$Iters \leftarrow \lceil log_r(c) \rceil$; $i \leftarrow 0$
$f \leftarrow$ pre-pruning network with parameter set $\boldsymbol{\theta}$
while $i < Iters$ **do**
  for $\boldsymbol{\theta}_{layer}$ **in** $\boldsymbol{\theta}$ **do**
    $\boldsymbol{\theta}'_{layer} \leftarrow \emptyset$
    while $|\boldsymbol{\theta}'_{layer}| < ||\boldsymbol{\theta}_{layer}||_0 \times (1-r)$ **do**
      $\tilde{k} \leftarrow \min_k(\Delta P_{k,\boldsymbol{\theta}'_{layer}})$ ▷ Find the greatest importance loss
      $\tilde{w} \leftarrow \text{argmax}_{w \in \boldsymbol{\theta}'_{layer}}(I_{\tilde{k},w})$ ▷ Find the highest importance for $\tilde{k}$
      $\boldsymbol{\theta}'_{layer} \leftarrow \boldsymbol{\theta}'_{layer} \bigcup \{\tilde{w}\}$
      $\Delta P_{k,\boldsymbol{\theta}'_{layer}} \leftarrow \frac{P_{k,\boldsymbol{\theta}'_{layer}} - P_{k,\boldsymbol{\theta}_{layer}}}{P_{k,\boldsymbol{\theta}_{layer}}}, \forall k \in K$ ▷ Update importance losses
    end while
    $\boldsymbol{\theta}_{prune} \leftarrow \{w | w \in \boldsymbol{\theta}_{layer} \wedge w \notin \boldsymbol{\theta}'_{layer}\}$
    $\boldsymbol{\theta}_{prune} \leftarrow \mathbf{0}$ ▷ Prune weights that are not selected
  end for
  $f \leftarrow$ Train $(\boldsymbol{\theta}; D)$; $i \leftarrow i + 1$
end while

Maintaining Share of Importance Based on the group importance scores, we compute the share of the importance of group k as follows:

$$P_k = \frac{I_k}{\sum_K I_k} \tag{7}$$

The share of importance in the original model $\boldsymbol{f}$ is used as a target. In the pruned model $\boldsymbol{f'}$ with parameter set $\boldsymbol{\theta'}$, the percentage change in the importance score compared to full model $\boldsymbol{f}$ is:

$$\Delta P_{k,\boldsymbol{f'}} = \frac{P_{k,\boldsymbol{\theta'}} - P_{k,\boldsymbol{\theta}}}{P_{k,\boldsymbol{\theta}}} \tag{8}$$

As weights are pruned, the importance scores for each group would inevitably decrease. However, the disparate loss of importance across groups leads to an imbalanced loss in classification performance. Thus, we apply a layer-wise greedy algorithm to select the parameters that minimize the difference of $\Delta P_{k,\boldsymbol{f}}$ between the sensitive groups, as explained in Algorithm 1.

FairGRAPE iteratively prunes and fine-tunes a network: given a desired sparsity c and a step size r, the percentage of remaining weights to be pruned at each iteration, the total number of iterations is $Iters = \lceil (log_r(c) \rceil$. In each iteration, the network is pruned layer by layer for all layers with weight attributes (*e.g.* convolutional and linear layers). Before pruning a layer, $P_{k,\boldsymbol{\theta}}$ for each group k is calculated with all unpruned parameters $\boldsymbol{\theta}$. At the very beginning of the algorithm, $\boldsymbol{\theta'}$ has not included any weights yet, and all group importance values are 0. So $P_{k,\boldsymbol{\theta'}}$ are initialized to $1/|K|$. Then weights in $\boldsymbol{\theta}$ are added to $\boldsymbol{\theta'}$ one

Dataset	Images	White	Black	Hispanic	East Asian	Southeast Asian	Indian	Middle Eastern	Male	Female	Categories
FairFace [30]	97,698	18,612	13,789	14,990	13,837	12,210	13,835	10,425	51,778	45,920	-
UTKFace [63]	22,013	10,078	4,526	-	3,434	-	3,975	-	11,631	10,382	-
CelebA [38]	202,599	-	-	-	-	-	-	-	84,434	118,165	39
ImageNet(Person) [56]	10,215	-	-	-	-	-	-	-	6,590	3,625	103

Table 1: Demographic composition of datasets.

at a time. Before each selection, the sensitive group k with the minimum $\Delta P_{k,\boldsymbol{\theta}'}$ is identified, as shown in line 9 of Algorithm 1. Then the weight that has the highest importance score for group k is added to the set of selected weights to minimize $P_{k,\boldsymbol{\theta}'}$ (line 10). $P_{k,\boldsymbol{\theta}'}$ and $\Delta P_{k,\boldsymbol{\theta}'}$ are updated for all groups $k \in K$ (line 12). The selection for weights continues until $(1-r)\%$ of weights are selected. The weights not selected are removed by setting them to zero and thus no longer considered in further iterations. Then FairGRAPE proceeds to the next layer. Once all layers are pruned, the network is retrained for a fixed number of e epochs to adjust the weights to its current structure. Then the next iteration begins.

4 Experiments

4.1 Datasets

To evaluate our proposed FairGRAPE, we conducted extensive experiments with four face image datasets, including FairFace [30], UTKFace [63], CelebA [38], and the person subtree of ImageNet [56]. Table 1 shows the distributions of races and genders in all datasets. Images are fairly distributed across the seven race groups in the FairFace, while the white race is dominant in the UTKFace. This allows us to validate that the effect of our method remains consistent with the presence of data bias. In UTKFace, only one "*Asian*" contains both Asian and Southeast Asian faces. We excluded the "*Other*" category in UTKFace due to its ambiguity. Race/ethnicity information is not provided in CelebA and ImageNet. FairFace, UTKFace, and CelebA provide annotations for binary genders. While the ImageNet person subtree contains three gender classes: Male, Female, and Unsure (non-binary), we only use ImageNet samples with binary genders to stay consistent with other datasets. Following the practice in [56], Imagenet samples that are from "unsafe" categories or have imageability scores ≤ 4 are also excluded.

4.2 Experiment Settings

Network architectures: To ensure our method applies to different architectures, we use two popular deep networks: ResNet-34 [22] and MobileNet-V2 [27]. ResNet is widely applied for classification tasks, and the MobileNet is a compact network commonly used by mobile devices. All models are pre-trained on ImageNet [9].

Hyperparameters: We use a cross-entropy loss function with the ADAM optimizer for all training. All accuracy scores, overall and group-wise, are averaged across three trials to control for randomness in training. For iterative pruning methods, we retrain five epochs after each pruning iteration. Step size $r = 0.9$ on FairFace, CelebA and Imagenet and $r = 0.975$ on UTKFace. The training/validation/testing percentage is 80%/10%/10% in each dataset.

4.3 Baseline Methods

We deploy the following four baseline methods: **Single-Shot Network Pruning (SNIP)** [34]: calculates the connection sensitivity of edges by back-propagating on one mini-batch and prunes the edges with low sensitivity. **Weight Selection (WS)** [19]: prunes the weights with magnitudes below a threshold in a trained model. It is the most commonly used in mobile applications [42]. **Lottery Ticket Identification (Lottery)** [13]: records the initial state of the network; resets the model to its initial state after each pruning iteration. **Gradient Signal Preservation (GraSP)** [50]: removes the parameters with low Hessian-gradient scores to maximize gradient signal in the pruned model.

5 Results

To evaluate the effectiveness of our method, we conduct extensive experiments on three different settings, including (§5.1) gender and race classification tasks, (§5.2) non-sensitive attribute classification tasks, and (§5.3) model pruning based on unsupervised clustering. We also perform more in-depth analysis, including (§5.4) ablation studies, (§5.6) different sparsity levels, (§5.5) pruning on minority faces, and (§5.7) difference in importance score and structure, to understand the importance of components in FairGRAPE.

5.1 Gender and Race Classification

We first perform experiments to verify bias mitigation in classifying sensitive attributes. Table 2 shows classification accuracy and biases on FairFace and UTKFace datasets where we compress the ResNet-34 and MobileNet-V2, respectively. The column 'Task' indicates the dataset and classification task. We report overall classification accuracy, accuracy by sensitive groups, and variances in accuracy degradation. FairGRAPE consistently produces a substantially higher accuracy, lower differences in accuracy, and lower variance in performance degradation than the baseline methods. For example, SNIP sometimes produces accuracy scores close to our method, but it has a remarkably larger accuracy variance than FairGRAPE, which implies the potential biases caused by model pruning. In the only cases of FairFace, WS produced a model with a smaller race classification accuracy gap between male and female images, but at the cost of drastically worsened accuracy for both groups. These results suggest that our proposed method successfully equalizes the impact of pruning on the sensitive

Task	Method	Accuracy			Bias		Accuracy								Bias	
		All	Male	Female	$\rho(A)$	$\rho(\Delta)$	All	White	Black	Hisp	E-A	SE-A	Indian	ME	$\rho(A)$	$\rho(\Delta)$
FairFace, Gender	No-pruning	94.6	94.7	94.5	0.14	-	94.6	94.6	90.5	95.9	94.7	94.4	96.3	95.6	1.93	-
	Lottery	85.8	86.4	85.2	0.80	0.65	85.8	85.1	80.8	88.4	84.0	85.5	88.1	89.6	3.01	1.55
	SNIP	90.4	91.0	89.9	0.78	0.63	90.4	91.0	85.2	92.6	90.0	90.5	91.3	92.6	2.53	0.93
	WS	83.8	84.3	83.4	0.62	0.47	83.9	82.9	78.9	87.2	82.2	82.2	86.2	88.3	3.32	2.00
	GraSP	87.9	88.4	87.4	0.75	0.60	87.9	87.5	83.1	89.6	87.5	88.0	89.4	90.9	2.49	0.93
	FairGRAPE	**91.1**	91.3	91.0	**0.20**	**0.05**	**90.5**	90.4	85.4	92.3	90.1	90.5	91.9	92.8	**2.47**	**0.77**
FairFace, Race	No-pruning	72.0	71.2	72.9	1.23	-	72.0	73.9	83.2	59.6	77.6	66.9	75.4	66.2	8.02	-
	Lottery	57.1	55.3	59.1	2.64	1.42	57.1	69.7	78.8	33.0	74.1	43.5	61.7	30.4	20.0	12.9
	SNIP	62.3	60.4	64.3	2.78	1.55	62.3	74.1	80.8	44.5	73.7	53.7	66.0	34.8	17.1	10.7
	WS	47.9	47.3	48.5	**0.86**	**0.36**	47.9	64.7	77.9	8.61	78.3	31.1	37.8	30.0	26.9	19.9
	GraSP	57.9	56.0	60.1	2.88	1.55	57.9	69.6	77.3	38.6	72.0	47.0	62.1	30.7	18.0	11.3
	FairGRAPE	**66.8**	65.3	68.6	2.35	1.12	**65.1**	72.2	80.3	47.5	75.8	56.3	70.2	48.6	**13.4**	**6.13**
UTKFace, Gender	No-pruning	93.5	92.4	94.8	1.68	-	93.5	94.1	-	95.1	-	89.6	-	93.7	2.45	-
	Lottery	83.5	83.7	83.3	0.34	2.01	83.5	84.7	-	85.8	-	75.0	-	85.2	5.15	2.79
	SNIP	91.0	91.3	90.6	0.45	2.19	91.0	91.9	-	93.0	-	86.0	-	90.9	3.08	0.67
	WS	81.9	81.4	82.6	0.89	1.79	81.9	82.1	-	84.9	-	77.2	-	82.4	3.20	0.92
	GraSP	86.8	88.5	84.9	2.51	4.20	86.8	86.7	-	89.8	-	81.4	-	88.3	3.66	1.43
	FairGRAPE	**92.2**	92.0	92.5	**0.31**	**1.36**	**91.9**	92.7	-	94.0	-	87.9	-	91.3	**2.61**	**0.56**
UTKFace, Race	No-pruning	90.8	90.6	90.9	0.24	-	90.8	92.2	-	92.5	-	93.3	-	83.3	4.69	-
	Lottery	71.7	69.4	74.2	3.41	3.17	71.7	83.8	-	80.3	-	61.0	-	42.7	19.0	15.6
	SNIP	86.8	85.7	88.0	1.64	1.40	86.8	91.6	-	92.5	-	85.8	-	70.1	10.4	6.28
	WS	70.7	68.3	73.5	3.68	3.41	70.7	82.7	-	80.8	-	59.2	-	41.4	19.6	16.2
	GraSP	77.7	76.4	79.1	1.94	1.70	77.7	86.1	-	83.3	-	72.2	-	56.4	13.5	9.81
	FairGRAPE	**88.7**	88.2	89.3	**0.78**	**0.54**	**88.5**	90.6	-	92.2	-	88.9	-	79.0	**5.93**	**2.04**

Table 2: The group-wise accuracy and biases in gender or race classification tasks. Hisp, E-A, SE-A and ME stand for Hispanic, East Asian, Southeast Asian and Middle Eastern. FairFace experiments are conducted on ResNet-34 pruned at 99% sparsity, UTKFace experiment on MobileNet-V2 pruned at 90% sparsity. $\rho(A)$ and $\rho(\Delta)$ are the standard deviation of accuracy and accuracy loss across sensitive groups, respectively.

groups regardless of the classification task, thus achieving a better trade-off between fairness and overall accuracy.

Furthermore, FairGRAPE shows solid performances in all settings with different architectures and datasets (balanced or imbalanced), proving the proposed method's robustness. See supplementary material for results when we jointly control race and gender groups.

We next visualize the proportion changes of False Negative Rates(FNRs)/False Positive Rate(FPRs) from the full model after pruning by FairGRAPE and other baseline methods in Figure 3. Each point in the plot represents normalized FNR and FPR change of a specific race group in the model produced by one of the pruning methods, and the ellipses are created by estimating a 95% confidence region of data points. The results reveal that the proposed FairGRAPE produces data points closer to the origin than the other data points generated by the baseline methods. More importantly, FairGRAPE creates the smallest ellipse, which demonstrates that performance changes for each group are close to each other. Thus the distribution of induced bias across sensitive groups is fair.

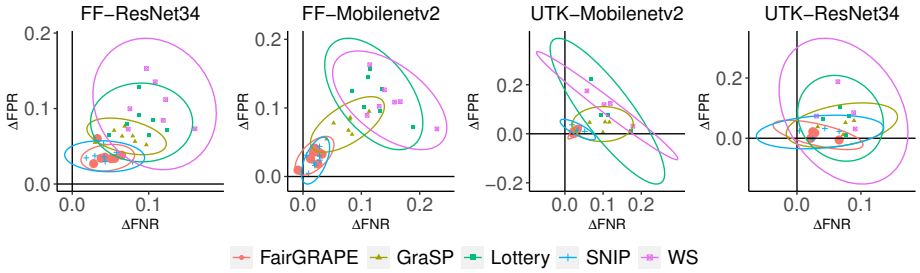

Fig. 3: Normalized FNR/FPR changes in race classification. Sparsity levels are 99% for ResNet-34 and 90% for MobileNet-V2. Each data point represents the mean value of a race. The ellipses are created by estimating 95% confidence ellipses, assuming multivariate t-distribution of points produced by each method.

5.2 Non-Sensitive Attribute Classification

To evaluate the performance of FairGRAPE in more practical cases where output classes and sensitive groups are disjoint, we experiment with classification on CelebA and ImageNet datasets. CelebA contains the 39 non-sensitive categories of facial attributes such as eyeglasses, makeup, and lipsticks. We code each of these categories as a binary classification task. For the ImageNet experiment, we use the modified person subtree, which contains 10,215 images in 103 distinct classes (*e.g.* , basketball player, rapper) with gender labels [56]. We train the models to classify the class to which a given image belongs. Note that we use the ResNet-34 network at 50% sparsity for the ImageNet experiments and the MobileNet-V2 network at 90% sparsity for the CelebA experiments.

Table 3 shows the overall accuracy, accuracy of each gender, and the standard deviation of accuracy change. FairGRAPE achieves the highest accuracy on ImageNet. Although GraSP has a smaller accuracy gender gap than our method, its overall accuracy and variance of performance degradation are drastically worse. In the CelebA experiment, FairGRAPE has a significantly lower variance in accuracy change than other methods while achieving the highest accuracy. The results demonstrate that FairGRAPE performs well on sensitive attribute classification tasks and non-sensitive attributes, thus widely applicable in various applications.

5.3 Unsupervised Learning for Group Aware in Model Pruning

In practice, labels for sensitive attributes may not always be available. Therefore, we further examine the performance of our method on a dataset without demographic group labels through unsupervised group discovery.

Table 4 shows the accuracy and bias of experiments on the FairFace dataset. In this test, FairGRAPE conducts pruning by calculating the importance score of parameters for clusters learned from unsupervised learning as sensitive groups. Then we evaluate accuracy and bias with actual race labels. We labeled the

Dataset	Task	Group	Methods	Accuracy			Bias	
				All	Male	Female	Diff	$\rho(\Delta)$
ImageNet	Person Subtree (103 classes)	Gender	No-Pruning	50.25	53.03	45.60	7.43	-
			Lottery	50.85	54.03	45.98	8.05	2.55
			SNIP	47.85	50.89	42.76	8.13	0.49
			WS	51.11	54.06	46.16	7.90	0.33
			GraSP	15.36	17.03	12.57	4.47	2.10
			FairGRAPE	**51.12**	54.01	46.16	**7.85**	**0.30**
CelebA	Non-sensitive Facial Attributes (39 classes)	Gender	No-Pruning	91.81	91.76	91.86	0.11	-
			Lottery	89.31	88.99	89.54	0.55	0.32
			SNIP	90.29	90.05	90.46	0.41	0.21
			WS	88.57	88.15	88.87	0.72	0.43
			GraSP	89.40	89.08	89.63	0.55	0.32
			FairGRAPE	**90.90**	90.74	91.01	**0.27**	**0.11**

Table 3: The average accuracy and biases in person category classification and facial attributes classification on ResNet-34 at 50% sparsity and MobileNet-V2 network at 90% sparsity, respectively. $\rho(\Delta)$ is the standard deviation of accuracy loss across genders.

Task	Methods	Accuracy								Bias	
		All	White	Black	Hisp	E-A	SE-A	Indian	ME	$\rho(A)$	$\rho(\Delta)$
FairFace, Race	No-Pruning	72.0	73.9	83.2	59.6	77.6	66.9	75.5	66.2	8.03	-
	Lottery	57.1	69.7	78.8	33.0	74.1	43.5	61.7	30.4	20.0	12.9
	SNIP	62.3	74.1	80.8	44.5	73.7	53.7	66.0	34.8	17.1	10.7
	WS	47.9	64.7	78.0	8.6	78.3	31.1	37.8	30.0	26.9	19.9
	GraSP	57.9	69.6	77.3	38.6	72.0	47.0	62.1	30.7	18.0	11.3
	FairGRAPE	**63.5**	69.7	80.4	49.2	74.7	53.7	68.1	42.8	**14.1**	**7.40**

Table 4: The average accuracy and biases in race classification, where FairGRAPE pruning is performed based on groups clustered by unsupervised learning. Hisp, E-A, SE-A and ME stand for Hispanic, East Asian, Southeast Asian and Middle Eastern. $\rho(A)$ and $\rho(\Delta)$ are the standard deviation of accuracy and accuracy loss across sensitive races.

seven clusters using K-means clustering on image embedding generated by the ResNet-34 network pre-trained on Imagenet. While all baseline methods have low accuracy and large variance of accuracy as they do not consider the sensitive groups, the FairGRAPE method consistently results in the lowest performance variance, suggesting that our proposed method has the potential to compress the model while reducing biases even in the absence of sensitive attribute information. The K-means algorithm's simplicity further reinforced our method's generalizability when the precise group partitioning is complex or noisy.

5.4 Ablation Studies: Group Importance and Iterative Pruning

Table 5 shows the performance of FairGRAPE with different group importance and iterative retraining settings. We first find that group importance is the essential component in our proposed method. The baseline method, which does

Group Importance	Iterative retraining (# iterations/r)	% Training Images	Accuracy			Bias	
			All	Female	Male	Diff	$\rho(\Delta)$
✓	✓ (22/0.1)	20%	90.90	91.01	90.74	0.27	0.11
✓	✓ (22/0.1)	100%	90.66	90.81	90.45	0.36	0.18
✓	✓ (22/0.1)	50%	90.72	90.84	90.54	0.30	0.13
✓	✓ (22/0.1)	10%	90.84	90.97	90.67	0.30	0.13
✓	✓ (16/0.2)	20%	90.49	90.62	90.32	0.30	0.20
✓	✓ (3/0.5)	20%	90.34	90.52	90.10	0.42	0.22
✓	✗ (1/0.9)	20%	89.26	89.51	88.92	0.41	0.34
✗	✓ (22/0.1)	-	89.31	89.54	88.99	0.45	0.31
✗	✗ (1/0.9)	-	88.57	88.86	88.17	0.69	0.42

Table 5: The accuracy and biases under different pruning settings. The MobileNet-V2 networks are trained on CelebA attributes classification tasks and pruned at 90% sparsity. # iter is the number of pruning iterations, determined by the pruning step r which is the proportion of remain edges removed during each iteration. % training images represents the percentage of training images included in calculation of group importance scores. $\rho(\Delta)$ is the standard deviation of accuracy loss across genders.

not use both group importance and iterative retraining, has remarkably lower accuracy, gender gap, and variance of accuracy changes than our method, which utilizes both components. As the pruning step r at each iteration increased, the accuracy decreased, and the bias increased gradually.

More specifically, the model suffers from an obvious performance drop and bias increase when r increased from 0.1 to 0.2. This result agrees with previous findings [13] that iterative pruning improves performance.

Finally, we examine the percentage of training images used in importance calculation. FairGRAPE calculates group-wise importance score $I_{w,k} = (g_w w)^2$ for each weight w, where g_w is calculated with respect to average loss across selected mini-batches of the training set. It has been found that the proportion of training images used in the calculation process affects pruning speed and accuracy [41]. We compared the performance using 100%, 50%, 20%, and 10% of training sets. The result indicates that 20% is the ideal ratio that produces the best performance.

5.5 Pruning on Images from Minority Races

This section examines whether rebalancing the dataset could mitigate pruning-induced bias. Using the UTKFace dataset, where white faces are dominant, we tested SNIP and GraSP with their gradient calculation and parameter selection conducted on non-white examples only (i.e., Black, Asian, Indian) Table 6 shows the result. Interestingly, using a subset of data did not significantly change overall accuracy. However, the overall biases increased compared to the case of using all data. This change shows that the problem of biases in pruned methods cannot

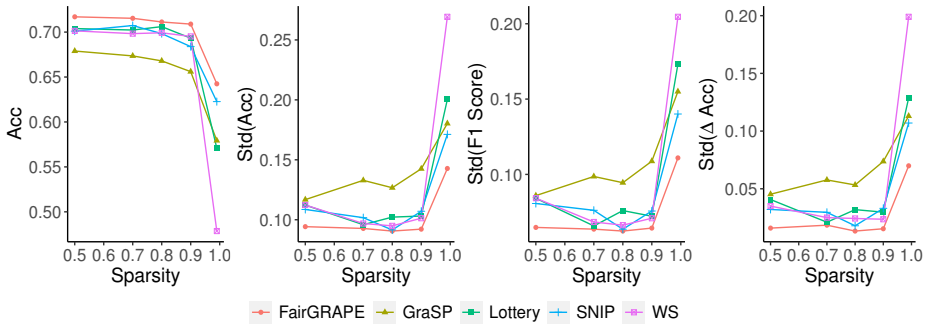

Fig. 4: Accuracy and biases of race classification across races at different sparsity levels. Experiments were conducted using ResNet-34 on FairFace dataset.

be solved by simple data rebalancing and our method effectively addresses this challenging problem.

Methods	Accuracy					Bias	
	All	White	Black	Asian	Indian	$\rho(A)$	$\rho(\Delta)$
No-pruning	93.84	95.08	95.27	89.85	92.70	2.54	-
FairGRAPE	91.72	92.86	94.18	86.88	90.40	3.21	0.78
GraSP (Minority)	89.15	88.73	92.24	83.71	91.19	3.80	2.31
GraSP (All data)	88.33	88.80	91.05	82.47	89.13	3.73	1.77
SNIP (Minority)	90.55	91.60	92.79	83.91	91.19	4.03	1.90
SNIP (All data)	90.95	91.33	94.18	85.44	91.11	3.66	1.62

Table 6: UTKFace gender classification accuracy on minority subsets. $\rho(A)$ and $\rho(\Delta)$ are the standard deviation of accuracy and accuracy loss across racs, respectively.

5.6 Analysis on Model Sparsity Levels

We next evaluate the performance of FairGRAPE across different sparsity levels to understand its effectiveness. Figure 4 shows changes in accuracy and biases over different sparsity levels. FairGRAPE outperforms the baseline methods by producing the highest accuracy and lowest disparity of performance degradation across sensitive groups at various pruning rates. As sparsity changes from 90% to 99%, most baseline methods exhibit a sharp decrease in accuracy and increase in bias, while performance change in FairGRAPE is substantially smaller. This confirms that our method can be widely deployed to real-world systems with various sparsity levels.

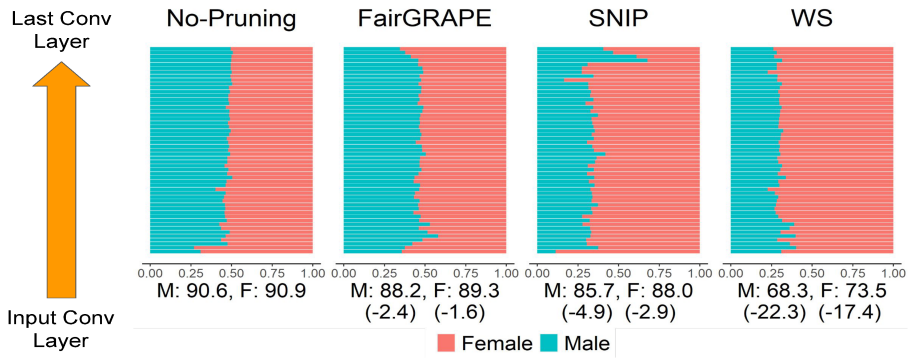

Fig. 5: The ratio of importance scores on MobileNet-V2. Networks are pruned to 90% sparsity and trained on UTKFace dataset. M and F indicate race classification accuracy on male and female images. Accuracy changes between the pruned models and the full model are shown in parenthesis.

5.7 Layer-wise Importance Scores and Bias

This subsection performs an in-depth structural analysis on pruned networks. Figure 5 visualizes the ratio of importance scores at each layer for each gender group. Each bar represents a convolutional or linear layer. and the width of a colored segment indicates the ratio of importance score for the corresponding gender group. FairGRAPE preserves the balanced importance distribution of the full network, with similar scores for both genders, leading to substantially smaller gaps in accuracy and accuracy change. The group-agnostic pruning methods, including SNIP and Weight Selection, select weights with higher importance for the female group, which is already showing higher accuracy in the original model. Consequently, the accuracy of the male group suffered from a substantially greater loss and the gap is much larger than the model pruned by FairGRAPE.

6 Conclusion

In this paper, we proposed FairGRAPE, a novel pruning method that prunes weights based on their importance with respect to each demographic sub-group in the dataset. Empirical results show that our method can minimize performance degradation across sub-groups in different network architectures and datasets at various pruning rates. We also demonstrated that the association between distributions of gradient importance and performance biases has an important implication for understanding information loss during model compression. Our work will therefore contribute to developing fair light-weight models that can be deployed on many mobile devices by mitigating hidden biases.

Acknowledgement This work was supported by NSF SBE-SMA #1831848.

References

1. Alvi, M., Zisserman, A., Nellåker, C.: Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2018) 3
2. Bahng, H., Chun, S., Yun, S., Choo, J., Oh, S.J.: Learning de-biased representations with biased representations. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 119, pp. 528–539 (2020) 3
3. Blakeney, C., Huish, N., Yan, Y., Zong, Z.: Simon says: Evaluating and mitigating bias in pruned neural networks with knowledge distillation. arXiv preprint arXiv:2106.07849 (2021) 1, 4
4. Blalock, D., Gonzalez Ortiz, J.J., Frankle, J., Guttag, J.: What is the state of neural network pruning? Proceedings of machine learning and systems **2**, 129–146 (2020) 1, 3
5. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FACCT). pp. 77–91 (2018) 3
6. Chen, Y., Joo, J.: Understanding and mitigating annotation bias in facial expression recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14980–14991 (2021) 3
7. Cheng, Y., Wang, D., Zhou, P., Zhang, T.: A survey of model compression and acceleration for deep neural networks. arXiv preprint arXiv:1710.09282 (2017) 1
8. Das, A., Dantcheva, A., Bremond, F.: Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2018) 3
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition (CVPR). pp. 248–255. Ieee (2009) 7
10. Du, M., Yang, F., Zou, N., Hu, X.: Fairness in deep learning: A computational perspective. IEEE Intelligent Systems (2020) 3
11. Dubey, A., Chatterjee, M., Ahuja, N.: Coreset-based neural network compression. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 454–470 (2018) 3
12. Dwork, C., Immorlica, N., Kalai, A.T., Leiserson, M.: Decoupled classifiers for group-fair and efficient machine learning. In: Conference on fairness, accountability and transparency (FAACT). pp. 119–133. PMLR (2018) 3
13. Frankle, J., Carbin, M.: The lottery ticket hypothesis: Finding sparse, trainable neural networks. In: Proceedings of the International Conference on Learning Representations (ICLR) (2019) 1, 3, 8, 12
14. Frankle, J., Dziugaite, G.K., Roy, D.M., Carbin, M.: Stabilizing the lottery ticket hypothesis. arXiv preprint arXiv:1903.01611 (2019) 1
15. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Synthetic data augmentation using gan for improved liver lesion classification. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI). pp. 289–293 (2018) 3
16. Garcia, R.V., Wandzik, L., Grabner, L., Krueger, J.: The harms of demographic bias in deep face recognition research. In: 2019 International Conference on Biometrics (ICB). pp. 1–6 (2019) 3

17. Georgopoulos, M., Panagakis, Y., Pantic, M.: Investigating bias in deep face analysis: The kanface dataset and empirical study. Image and Vision Computing **102**, 103954 (2020) 3
18. Gwilliam, M., Hegde, S., Tinubu, L., Hanson, A.: Rethinking common assumptions to mitigate racial bias in face recognition datasets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 4123–4132 (2021) 3
19. Han, S., Mao, H., Dally, W.J.: Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In: Proceedings of the International Conference on Learning Representations (ICLR) (2016) 1, 3, 8
20. Han, S., Pool, J., Tran, J., Dally, W.: Learning both weights and connections for efficient neural network. In: Advances in Neural Information Processing Systems. vol. 28 (2015) 3
21. Hazirbas, C., Bitton, J., Dolhansky, B., Pan, J., Gordo, A., Ferrer, C.C.: Towards measuring fairness in ai: the casual conversations dataset. arXiv preprint arXiv:2104.02821 (2021) 3
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 7
23. He, Y., Zhang, X., Sun, J.: Channel pruning for accelerating very deep neural networks. In: Proceedings of the IEEE international conference on computer vision. pp. 1389–1397 (2017) 3
24. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. NIPS Deep Learning and Representation Learning Workshop (2015) 3
25. Hooker, S., Courville, A., Clark, G., Dauphin, Y., Frome, A.: What do compressed deep neural networks forget? arXiv preprint arXiv:1911.05248 (2019) 1, 4
26. Hooker, S., Moorosi, N., Clark, G., Bengio, S., Denton, E.: Characterising bias in compressed models (2020), arXiv preprint arXiv:2010.03058 1, 4
27. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017) 7
28. Joo, J., Kärkkäinen, K.: Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation. In: Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia. pp. 1–5 (2020) 3
29. Jung, S., Chun, S., Moon, T.: Learning fair classifiers with partially annotated group labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10348–10357 (2022) 3
30. Karkkainen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1548–1558 (2021) 1, 2, 3, 7
31. Kleindessner, M., Samadi, S., Awasthi, P., Morgenstern, J.: Guarantees for spectral clustering with fairness constraints. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning (ICML). vol. 97, pp. 3458–3467 (2019) 3
32. Krishnan, A., Almadan, A., Rattani, A.: Understanding fairness of gender classification algorithms across gender-race groups. In: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 1028–1035. IEEE (2020) 3

33. Lee, J., Kim, E., Lee, J., Lee, J., Choo, J.: Learning debiased representation via disentangled feature augmentation. Advances in Neural Information Processing Systems (NIPS) **34** (2021) 3
34. Lee, N., Ajanthan, T., Torr, P.: Snip: Single-shot network pruning based on connection sensitivity. In: Proceedings of the International Conference on Learning Representations (ICLR) (2019) 1, 3, 8
35. Li, H., Kadav, A., Durdanovic, I., Samet, H., Graf, H.P.: Pruning filters for efficient convnets. International Conference on Learning Representations (ICLR) (2017) 3
36. Li, Y., Vasconcelos, N.: Repair: Removing representation bias by dataset resampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3
37. Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., Zhang, C.: Learning efficient convolutional networks through network slimming. In: Proceedings of the IEEE international conference on computer vision. pp. 2736–2744 (2017) 3
38. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015) 2, 7
39. Misra, I., Zitnick, C.L., Mitchell, M., Girshick, R.: Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 3
40. Molchanov, P., Tyree, S., Karras, T., Aila, T., Kautz, J.: Pruning convolutional neural networks for resource efficient inference. In: International Conference on Learning Representations (ICLR) (2017) 3
41. Molchanov, P., Mallya, A., Tyree, S., Frosio, I., Kautz, J.: Importance estimation for neural network pruning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11264–11272 (2019) 3, 5, 12
42. Nan, K., Liu, S., Du, J., Liu, H.: Deep model compression for mobile platforms: A survey. Tsinghua Science and Technology **24**(6), 677–693 (2019) 8
43. Ramaswamy, V.V., Kim, S.S.Y., Russakovsky, O.: Fair attribute classification through latent space de-biasing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9301–9310 (2021) 3
44. Ryu, H.J., Adam, H., Mitchell, M.: Inclusivefacenet: Improving face attribute detection with race and gender diversity. arXiv preprint arXiv:1712.00193 (2017) 3
45. Schumann, C., Wang, X., Beutel, A., Chen, J., Qian, H., Chi, E.H.: Transfer of machine learning fairness across domains. Clinical Orthopaedics and Related Research (CoRR) (2019) 3
46. Stoychev, S., Gunes, H.: The effect of model compression on fairness in facial expression recognition. arXiv preprint arXiv:2201.01709 (2022) 4
47. Tai, C., Xiao, T., Zhang, Y., Wang, X., Weinan, E.: Convolutional neural networks with low-rank regularization. In: International Conference on Learning Representations (ICLR) (2016) 3
48. Terhörst, P., Kolf, J.N., Damer, N., Kirchbuchner, F., Kuijper, A.: Face quality estimation and its correlation to demographic and non-demographic bias in face recognition. In: 2020 IEEE International Joint Conference on Biometrics (IJCB). pp. 1–11 (2020) 3
49. Wang, A., Barocas, S., Laird, K., Wallach, H.: Measuring representational harms in image captioning. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 324–335 (2022) 3

50. Wang, C., Zhang, G., Grosse, R.: Picking winning tickets before training by preserving gradient flow. In: International Conference on Learning Representations (ICLR) (2020) 1, 3, 8
51. Wang, M., Deng, W., Hu, J., Tao, X., Huang, Y.: Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 1, 3
52. Wang, T., Zhao, J., Yatskar, M., Chang, K.W., Ordonez, V.: Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 3
53. Wang, W., Fu, C., Guo, J., Cai, D., He, X.: Cop: customized deep model compression via regularized correlation-based filter-level pruning. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. pp. 3785–3791 (2019) 3
54. Wang, Z., Qinami, K., Karakozis, I.C., Genova, K., Nair, P., Hata, K., Russakovsky, O.: Towards fairness in visual recognition: Effective strategies for bias mitigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 3
55. Xu, X., Huang, Y., Shen, P., Li, S., Li, J., Huang, F., Li, Y., Cui, Z.: Consistent instance false positive improves fairness in face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 578–586 (2021) 3
56. Yang, K., Qinami, K., Fei-Fei, L., Deng, J., Russakovsky, O.: Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FACCT). pp. 547–558 (2020) 2, 7, 10
57. Yang, Y., Gupta, A., Feng, J., Wu, Y., Yadav, V., Hedau, V., Singhal, P., Natarajan, P., Joo, J.: Explaining deep convolutional neural networks via latent visual-semantic filter attention. In: 5th AAAI/ACM Conference on AI, Ethics, and Society (2022) 3
58. Yang, Y., Kim, S., Joo, J.: Explaining deep convolutional neural networks via latent visual-semantic filter attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8333–8343 (2022) 3
59. You, H., Li, C., Xu, P., Fu, Y., Wang, Y., Chen, X., Baraniuk, R.G., Wang, Z., Lin, Y.: Drawing early-bird tickets: Toward more efficient training of deep networks. In: International Conference on Learning Representations (ICLR) (2020) 3
60. Yu, R., Li, A., Chen, C.F., Lai, J.H., Morariu, V.I., Han, X., Gao, M., Lin, C.Y., Davis, L.S.: Nisp: Pruning networks using neuron importance score propagation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9194–9203 (2018) 3
61. Zafar, M.B., Valera, I., Rogriguez, M.G., Gummadi, K.P.: Fairness constraints: Mechanisms for fair classification. In: Artificial Intelligence and Statistics. pp. 962–970. PMLR (2017) 3
62. Zhang, C., Patras, P., Haddadi, H.: Deep learning in mobile and wireless networking: A survey. IEEE Communications surveys & tutorials **21**(3), 2224–2287 (2019) 1
63. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2017) 2, 7

64. Zhao, D., Wang, A., Russakovsky, O.: Understanding and evaluating racial biases in image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14830–14840 (2021) 3
65. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2979–2989 (2017) 3