A Joint Controlling Multiple Sensitive Groups

In real-world applications, various sensitive attributes, such as gender, race, and age, are often implicitly related to model outputs. Since the experiments in our main paper have only focused on considering one sensitive group to compress the original models, in this section, we further evaluate the performance of pruned networks by controlling for multiple sensitive attributes jointly.

Table 1 shows the accuracy and bias on FairFace and UTKFace in three different tasks, including gender, race, and gender-race classification. Each gender-race group is one combination of different genders and races, for example, White females and Black males. We measure the overall accuracy as well as group-wise biases. Note that we use the same compressed model by using both gender and race as the sensitive groups for all three tasks instead of using a particular sensitive group to prune the networks for each task.

Compared with experiments where genders and races are considered separate sensitive groups, all methods have a slightly higher bias and lower accuracy. We find that FairGRAPE remains the one with the lowest bias and highest accuracy, confirming its ability to maintain fairness for different intersections of sensitive groups when multiple groups are present. This result suggests that controlling multiple sensitive groups for pruning can be beneficial as the compressed model can be applied to different downstream tasks instead of controlling one sensitive group for each task.

		FairFace ResNet-34, 90% sparsity				**UTKFace** MobileNet-V2, 90% sparsity			
Task	Methods	Acc(↑)	FNR(↓)	FPR(↓)	Std(*Δ* Acc)(↓)	Acc(↑)	FNR(↓)	FPR(↓)	Std(*Δ* Acc)(↓)
Gender	No-Pruning	94.2	5.81	5.78	-	93.7	7.29	5.89	-
	Lottery	85.5	13.7	15.1	2.18	83.5	8.46	7.37	4.18
	SNIP	88.7	12.6	10.1	1.95	89.0	5.29	6.10	1.64
	WS	80.9	15.9	22.0	4.68	73.3	14.4	13.0	13.0
	GraSP	84.6	14.8	15.9	2.00	83.1	8.51	8.66	3.80
	FairGRAPE	**91.0**	**10.3**	**7.81**	**1.83**	**89.3**	**5.28**	**5.03**	**1.40**
Race	No-Pruning	72.2	28.2	4.65	-	90.5	8.09	4.73	-
	Lottery	56.5	44.7	7.32	13.8	74.2	30.0	9.68	14.1
	SNIP	60.3	40.8	6.65	12.1	83.0	19.4	6.15	7.41
	WS	47.2	53.5	8.87	19.5	60.3	47.2	15.1	22.6
	GraSP	56.0	45.2	7.38	10.8	71.1	33.2	8.46	15.1
	FairGRAPE	**66.2**	**34.5**	**5.67**	**5.78**	**83.4**	**18.2**	**5.93**	**5.21**
Gender-Race	No-Pruning	68.4	32.5	2.44	-	84.7	15.9	2.23	-
	Lottery	48.6	53.3	3.97	10.9	62.0	41.6	5.67	11.3
	SNIP	53.9	48.1	3.56	10.0	74.2	28.2	3.80	6.69
	WS	38.4	63.2	4.77	17.0	44.8	61.2	8.29	19.7
	GraSP	48.2	53.5	4.00	8.22	59.4	44.7	6.06	11.8
	FairGRAPE	**60.8**	**40.5**	**3.03**	**5.35**	**74.9**	**26.7**	**3.21**	**6.05**

Table 1: The overall accuracy and biases in joint classification, where gender-race are sensitive groups. Gender-race groups are intersections of genders and races, *e.g.* White female.

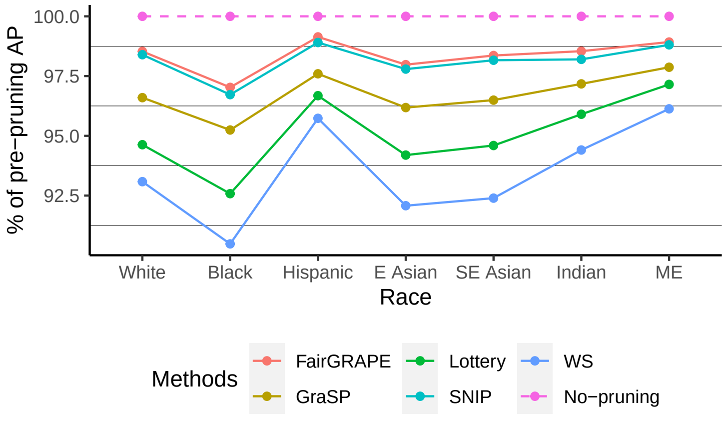


Fig. 1: Average Precision (AP) of each group in the gender classification task on FairFace. AP scores are represented as percentages of pre-pruning network value. The horizontal dashed line indicates the performance of the full model. FairGRAPE produced higher AP consistently across different races.

B Evaluation with AP

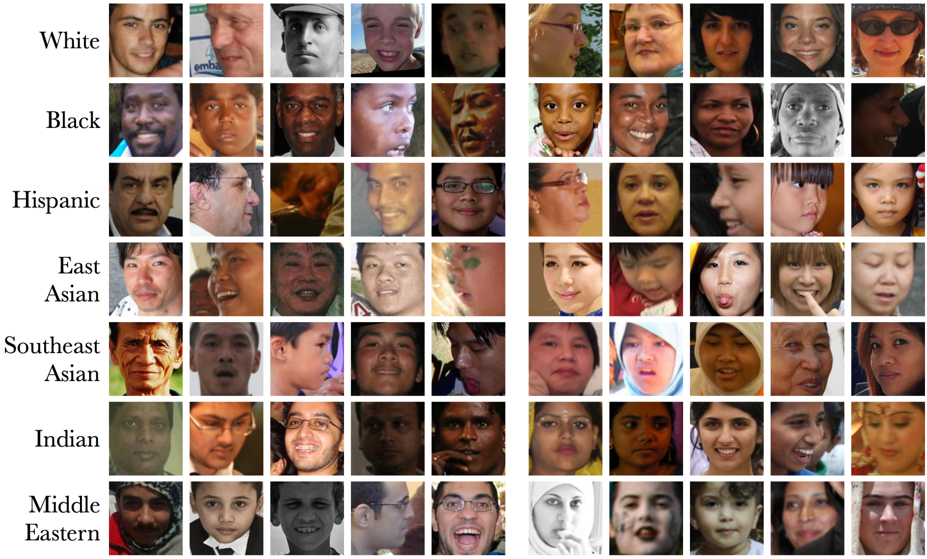
The main paper mainly used classification accuracy and false rates as the measurements because real-world classifiers must output a class rather than a continuous score. This section also reports experimental results using average precision (AP). Table 2 shows the mean and standard deviation of AP values by groups for race and gender classification on the FairFace dataset, as well as the standard deviation of differences between race AP values with corresponding full model AP values. Note that AP measures overall performances, the standard deviation of AP measures the performance gap across sensitive groups, and the standard deviation of AP differences measures the impact of pruning on the performance of each group. Although all groups suffer from performance degradation and an increase in disparity, FairGRAPE achieves the highest AP, the lowest AP disparity, and the lowest disparity in changes of AP.

C PIE (Pruning Identified Exemplars)

In this section, we perform a qualitative study to understand FairGRAPE's performance in preserving accuracy for individual samples. We first randomly sample face images from each race and gender group from FairFace and UTK-Face datasets, which provide annotation for both sensitive attribute. Figure 2 showcases the example face images. While most samples center at persons' faces without obstacles, they are taken at different angles and illuminations. Thus the level of visual challenges also significantly varies. In all groups, we find face examples that are not well lit, partially covered, not facing the camera, distorted,

Task	Group	Method	FairFace ResNet-34, 99% sparsity AP(↑)	Std(AP)(↓)	Std(ΔAP)(↓)	UTKFace MobileNet-V2, 90% sparsity AP(↑)	Std(AP)(↓)	Std(ΔAP)(↓)
Gender	Race	No-Pruning	0.988	0.009	-	0.981	0.009	-
		Lottery	0.940	0.23	0.015	0.912	0.024	0.015
		SNIP	0.970	0.017	0.010	0.961	0.018	0.010
		WS	0.924	0.027	0.020	0.901	0.018	0.010
		GraSP	0.956	0.017	0.009	0.941	0.025	0.018
		FairGRAPE	**0.971**	**0.016**	**0.007**	**0.959**	**0.017**	**0.009**
Race	Gender	No-Pruning	0.949	0.006	-	0.984	0.002	-
		Lottery	0.887	0.010	0.004	0.895	0.025	0.023
		SNIP	0.914	0.010	0.004	0.972	0.010	0.007
		WS	0.833	0.010	0.001	0.894	0.023	0.021
		GraSP	0.890	0.011	0.005	0.933	0.013	0.010
		FairGRAPE	**0.932**	**0.009**	**0.003**	**0.957**	**0.009**	**0.006**

Table 2: AP value in in gender classification, with races as sensitive groups. FairGRAPE produce the highest AP, the smallest variance between groups as well as the least disparity in changes.



(a) Male (b) Female

Fig. 2: Random samples from different gender-race groups.

or showing semantic attributes related to other groups. As studied in [2], such examples tend to receive greater impact from pruning.

We further investigate the impact of pruning over such samples and different sensitive groups using PIE (Pruning Identified Exemplars) [1, 2]. PIE refers to examples classified correctly by the full network while incorrectly by a pruned network. Figure 3 compares PIE and non PIE images. PIEs shown are random samples of images incorrectly classified by all pruning methods in race classifica-

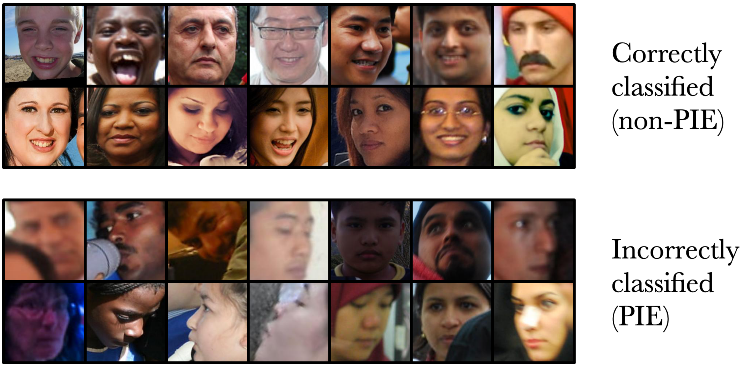


Fig. 3: Random samples of PIE and non-PIE face images.

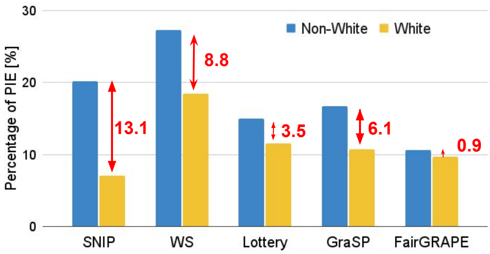


Fig. 4: PIE percentage difference between White and Non-White race groups of FairGRAPE and the baseline methods. Our method has a very small PIE difference, whereas other methods have large differences in PIE percentage between races.

tion. We find that the PIE examples identified frequently demonstrate visually challenging features shown in figure 2. While non-PIE faces are mostly clear, well-lit, taken from the front, and fully shown in the scope, PIE samples are often blurred, hard to distinguish, and show features from different race groups. This result demonstrates that visually challenging examples are more likely to suffer from pruning-induced bias.

We evaluate the percentage of such misclassified faces, specifically in White and Non-White race groups, of our method compared with the baseline methods as shown in Figure 4. Note that a high PIE percentage indicates that a large portion of misclassified faces. The result shows that FairGRAPE has no significant difference in PIE percentage between White and Non-White groups, while the baseline methods demonstrate extremely large differences. This result illustrates the capability of FairGRAPE to control for bias on the most challenging face exemplars.

References

1. Hooker, S., Courville, A., Clark, G., Dauphin, Y., Frome, A.: What do compressed deep neural networks forget? arXiv preprint arXiv:1911.05248 (2019) 3
2. Hooker, S., Moorosi, N., Clark, G., Bengio, S., Denton, E.: Characterising bias in compressed models (2020), arXiv preprint arXiv:2010.03058 3