Supplemental Material: Exploiting Transferable Knowledge for Fairness-aware Image Classification

Sunhee Hwang*, Sungho Park*, Pilhyeon Lee*, Seogkyu Jeon, Dohyung Kim, and Hyeran Byun†

Department of Computer Science, Yonsei University, Seoul, Republic of Korea
{sunny16, qkrtjdgh18, lph1114, jone9312, dohkim02, hrbyun}@yonsei.ac.kr

In this supplemental material, we provide an additional experimental result in terms of demographic parity. We also present the details of the architectures for the proposed classification networks.

1 Additional Experiment: Demographic Parity

Demographic Parity is one of the fairness definitions, which requires the equal rates of the positive outcome between different protected attribute groups [1]. Formally, all the protected attribute groups have the same positive outcome rates for the target attribute as follows:

$$\mathcal{P}(\hat{Y}=1|p=0)=\mathcal{P}(\hat{Y}=1|p=1), \tag{1}$$

where p, and $\hat{Y} \in \{0,1\}$ denote the protected attribute and the prediction.

For the quantitative evaluation, we measure the *Demographic Parity* (DP) defined as follows:

$$\mathcal{D}P = |PR_{p=0} - PR_{p=1}|, \tag{2}$$

where PR and p denote Positive Rate (PR) and a binary protected attribute respectively.

Table 1. *DP* for attractiveness classification. Lower is better.

Methods	Positive Rate		DP
	Young=0	Young=1	
ResNet-18 [2]	19.82	72.1	52.28
AdvDe [3]	27.76	72.75	42.9
PALL [4, 5]	20.21	63.11	44.99
Ours	38.43	81.23	**42.8**

*Equal contributions
†Corresponding author

Table. 1 shows the results of on CelebA dataset when the target attribute and protected attribute are set to Attractiveness and Young. Compare to other models, we achieve the fairest (the lowest DP) result.

2 Network Details

We present architectures of the protected attribute and target attribute classifiers as shown in Table. 2 and Table. 3.

Table 2. Architecture configurations for the Protected Attribute Classifier.

Layer Name	Layers	Output Size
conv1	$7 \times 7, 64, stride = 2$	$64 \times 32 \times 32$
Max Pool	$3 \times 3, stride = 2$	$64 \times 32 \times 16$
Res-block1	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$64 \times 32 \times 16$
Res-block2	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$128 \times 8 \times 8$
Res-block3	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$256 \times 4 \times 4$
Res-block4	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	$512 \times 2 \times 2$
Average Pool	$32 \times 3 \times 2$	$512 \times 1 \times 1$
Flatten		512
Fully Connected 1-1, 1-2, 1-3	512×100	100
Fully Connected 2-1, 2-2, 2-3	100×100	100
Fully Connected 3-1 (Gender)	100×2	2
Fully Connected 3-2 (Age)	100×6	6
Fully Connected 3-3 (Race)	100×5	5

References

1. Edwards, H., Storkey, A.: Censoring representations with an adversary. International Conference on Learning Representations (2016)
2. 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. (2016) 770–778
3. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. AIES '18, New York, NY, USA, Association for Computing Machinery (2018) 335–340

Table 3. Architecture configurations for the Target Attribute Classifier.

Layer Name	Layers	Output Size
conv1	$7 \times 7, 64, stride = 2$	$64 \times 32 \times 32$
Max Pool	$3 \times 3, stride = 2$	$64 \times 32 \times 16$
Res-block1	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$64 \times 32 \times 16$
Res-block2	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$128 \times 8 \times 8$
Res-block3	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$256 \times 4 \times 4$
Res-block4	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	$512 \times 2 \times 2$
Average Pool	$32 \times 3 \times 2$	$512 \times 1 \times 1$
Flatten		512
Fully Connected 1	512×100	100
Fully Connected 2	100×100	100
Fully Connected 3	100×2	2

4. Kim, B., Kim, H., Kim, K., Kim, S., Kim, J.: Learning not to learn: Training deep neural networks with biased data. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019)
5. 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: The IEEE International Conference on Computer Vision (ICCV). (2019)