Fair-VPT: Fair Visual Prompt Tuning for Image Classification

Supplementary Material

Method	BAcc. (↑)	Eopp. (↓)	DP (↓)
ViT [1]	68.7	55.9	41.6
VPT [4]	75.0	36.4	32.1
VPT [4]+AT [8]	63.2	23.4	24.1
VPT [4]+FSCL+ [7]	66.5	**11.8**	20.6
Fair-VPT (Ours)	**76.3**	12.1	**12.0**

Table 1. **Experimental results for *Attractive* on CelebA**. Eopp. and DP denote equal opportunity [3] and demographic parity [2] . We set *Gender* the the sensitive attribute.

Method	BAcc. (↑)	EOpp. (↓)	DP (↓)
ViT [1]	61.3	42.3	30.6
VPT [4]	62.8	40.4	28.5
VPT [4]+AT [8]	57.3	34.3	23.7
VPT [4]+FSCL+ [7]	63.6	41.0	25.2
Fair-VPT (Ours)	**65.3**	**18.3**	**15.9**

Table 2. **Experimental results for *Big Nose* on CelebA**. We set *Gender* the sensitive attribute.

Method	BAcc. (↑)	Eopp. (↓)	DP (↓)
ViT [1]	88.4	9.3	13.5
VPT [4]	89.0	11.3	12.1
VPT [4]+AT [8]	88.9	8.1	11.6
VPT [4]+FSCL+ [7]	89.0	9.6	9.9
Fair-VPT (Ours)	**90.9**	**4.1**	**4.9**

Table 3. **Experimental results on UTKFace.** The target label and sensitive attribute are respectively set to *Race* and *Gender*.

Method	BAcc. (↑)	EOpp. (↓)	DP (↓)
ViT [1]	74.8	53.2	49.0
VPT [4]	76.0	42.0	46.3
VPT [4]+AT [8]	77.5	45.6	43.2
Fair-VPT (Ours)	**80.7**	**37.6**	**37.2**

Table 4. **Experimental results on bFFHQ.** The target label and sensitive attribute are set to *Age* and *Gender* respectively.

1. Comparison Results with Other Metrics

We provide the comparison results measured by demographic parity [2] and equal opportunity [3] in Table 1, 2 3, 4, and 5.

2. Incorporation into VPT-deep Variant

As mentioned in the main paper, the proposed method is fundamentally designed based on VPT-shallow. However,

Method	BAcc. (↑)	EOpp. (↓)	DP (↓)
ViT [1]	74.8	53.2	49.0
VPT [4]	76.0	42.0	46.3
VPT [4]+AT [8]	77.5	45.6	43.2
Fair-VPT (Ours)	**80.7**	**37.6**	**37.2**

Table 5. **Experimental results on Waterbirds**. The target label is highly biased to the background.

Method	CelebA			UTK Face	
	Acc.	BAcc.	EO	BAcc.	EO
ViT [1]	78.4	68.7	41.6	88.4	13.4
VPT [4]	81.7	75.0	32.1	89.0	12.6
VPT [4]+AT [8]	67.6	63.2	24.0	88.9	11.6
VPT [4]+FSCL+[7]	69.3	66.5	20.6	89.0	9.9
Fair-VPT	78.6	76.3	**12.0**	**90.9**	**4.9**
VPT (deep) [4]	**82.3**	75.6	31.7	90.4	11.1
Fair-VPT (deep)	79.1	**76.5**	14.2	90.5	6.7

Table 6. **Incoporation into VPT-deep**. We set *Attractive* and *Race* to the target labels on CelebA and UTKFace, respectively. We set *Gender* to the sensitive attribute on both datasets.

it can be simply applied to the VPT-deep by prepending the prompts into the input space of each transformer layer. In Table 6, the proposed methods significantly enhance fairness in both the variants (*i.e.,* Fair-VPT and Fair-VPT (deep)). They each demonstrate superior performance in terms of accuracy and equalized odds (EO) respectively.

3. Discussion on Training Time

In this section, we compare the training time of the proposed method with the baseline (*i.e.,* VPT [4]). Compared to the baseline, our method further requires the classification and projection heads, which are single fully connected layers, and the calculation of the masked self-attention. However, the overhead of these additional components is not significant, and the backbone network (*i.e.,* ViT [1]) remains frozen. Therefore, the training time of our method is comparable to the baseline. Empirically, ours shows a training time of approximately 1.29 times that of the baseline on CelebA.

4. Discussion on Hyper-parameters

The hyper-parameters M and α influence the trade-off between accuracy and fairness. When α is relatively larger compared to M, accuracy tends to be improved. Meanwhile, when M is relatively larger than α, fairness tends to be en-

hanced. In addition, the absolute size of M affects the overall performance. Therefore, we fix it in all models as 50 for fair comparison.

5. More Details for Implementation

For GRL [8] and FSCL+ [7], we set the ratio for gradient reversal and the temperature to 1 and 0.1 for all the experiments. For ours, we determine λ to be 0.1 on UTKFace [10], bFFHQ [5], and Waterbirds [9], and 1.0 on CelebA [6]. The initial learning rates are set to 0.1 on CelebA, UTKFace, bFFHQ, and 0.01 on Waterbird.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1

[2] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery. 1

[3] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc. 1

[4] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, 2022. 1

[5] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. In *Advances in Neural Information Processing Systems*, 2021. 2

[6] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 2

[7] Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Fair contrastive learning for facial attribute classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10389–10398, 2022. 1, 2

[8] Edward Raff and Jared Sylvester. Gradient reversal against discrimination: A fair neural network learning approach. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 189–198, 2018. 1, 2

[9] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. 2

[10] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2