Supplementary Material for "FairViT: Fair Vision Transformer via Adaptive Masking"

Bowei Tian[1], Ruijie Du[2], and Yanning Shen[2(✉)]

[1] Wuhan University, Wuhan, Hubei 430072, China
boweitian@whu.edu.cn
[2] University of California, Irvine, CA 92697, USA
{ruijied,yannings}@uci.edu

The supplemental material consists of this appendix. This appendix includes an illustration of interpretability study (Section A), more experimental results on the ablation study of α and γ (Section B), and some implementation details (Section C).

A Interpretability Study

A.1 Gradient Attention Rollout

The Gradient Attention Rollout (GAR) [1] aims to illustrate why the attention mechanism performs well in many scenarios in computer vision. GAR achieves interpretability by a heat map highlighting how much areas contribute to the output. Specifically, GAR is defined as

$$\mathcal{A}_l = \begin{cases} \mathbf{A}_l(\mathbf{x}) \frac{\partial \hat{y}}{\partial \mathbf{A}_l(\mathbf{x})} \mathcal{A}_{l-1}, & \text{if } l > 0, \\ \mathbf{A}_l(\mathbf{x}) \frac{\partial \hat{y}}{\partial \mathbf{A}_l(\mathbf{x})}, & \text{if } l = 0, \end{cases} \tag{A1}$$

where $\mathbf{A}$ is an abbreviation of Attn in Equation (3), and $\mathcal{A}_l$ denotes the GAR on the l_{th} layer of the transformer. To generate the heat map, we assign the value $\mathcal{A}_N^{0,i}$ to the i_{th} patch in the image, where $\mathcal{A}_N$ represents the GAR of the last layer, measuring the importance of each patch in the final prediction. Note that $\mathcal{A}_N$ is a matrix, and $\mathcal{A}_N^{0,i}$ corresponds to the element at the 0-th row and i-th column. The primary objective of GAR is to quantify the relative importance of each input patch within the attention mechanism, and it is particularly useful for analyzing the model behavior and explaining its decision-making process [1].

A.2 Additional Implementation

In this section, we implement the interpretability study into two additional scenarios, i.e. $\mathbf{Y}$: Expression, $\mathbf{S}$: Attraction and $\mathbf{Y}$: Expression, $\mathbf{S}$: Hair color, to evaluate the effectiveness of FairViT. The corresponding images are shown in Figure A1, and the outcomes align consistently with our observations in Section 5.4. The Vanilla method captures information relevant to sensitive attributes. In contrast, FairViT exhibits a tendency to extract information relevant to the target attributes, such as Expression in the first scenario and Attraction in the

Fig. A1: The extended interpretability study of FairViT .

second scenario. Furthermore, from the last two scenarios, despite the variations in the sensitive attribute, the heat map remains capable in capturing the target attribute.

B Ablation Study: Standard Deviation Qualification

Due to the limited space in Table 3 and 4, we present the impact of α and γ in a figure manner, adding the standard deviation to elaborate the systematic error in our experiments. The results are shown in Figure B2 and Figure B3.

Impact of α. As α surpasses the threshold of 1, a noticeable decline in accuracy is observed, coupled with an escalation in standard deviation. This suggests that a performance decline leads to a more fluctuating demonstration. Additionally, it is noteworthy that the results are not substantially sensitive to parameter selection for a reasonable range.

Impact of γ. Compared to α, γ maintains a comparatively low standard deviation and demonstrates less fluctuation across different values. There is a subtle performance peak observed at around $\gamma = 0.5$.

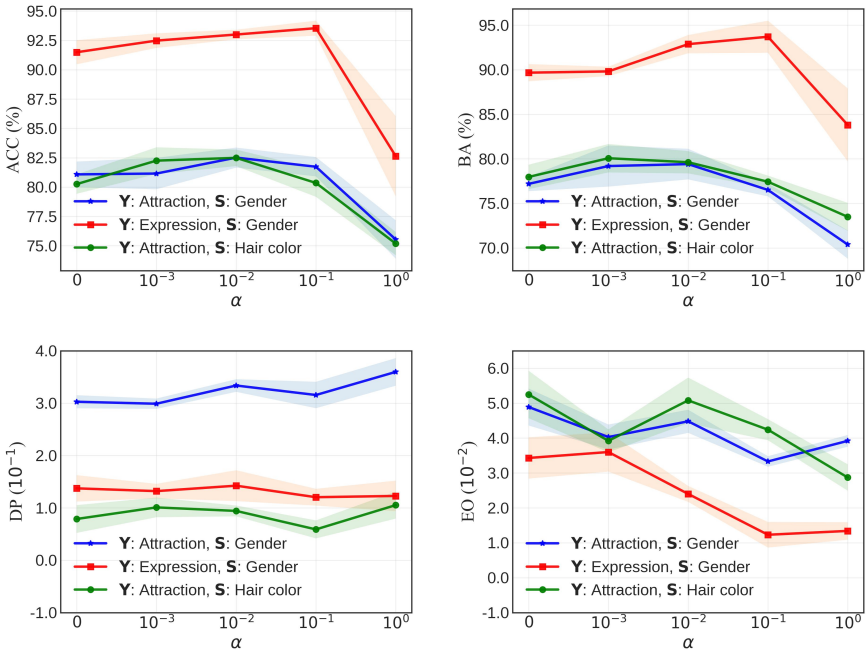

Fig. B2: Impact of α. Shown is the mean $\pm$ standard deviation of 3 runs with different random seeds.

C Implementation Details

The models are trained offline using PyTorch [2] and executed on a machine equipped with an AMD Ryzen Threadripper 3970X 32-Core CPU @ 2.00GHz and an NVIDIA GeForce RTX A4000 GPU, running the Ubuntu 20.04 operating system. To ensure a consistent data flow during training and to save computing power, we opt to use the first 80 individuals from the CelebA dataset [3] rather than the entire dataset.

To mitigate overfitting in the distance loss, we establish a lower bound of -2 for the calculation of each sample, further details are illustrated in the code. The code is available at https://github.com/abdd68/Fair-Vision-Transformer.

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. arXiv preprint arXiv:2005.00928 (2020)
2. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems 32 (2019)

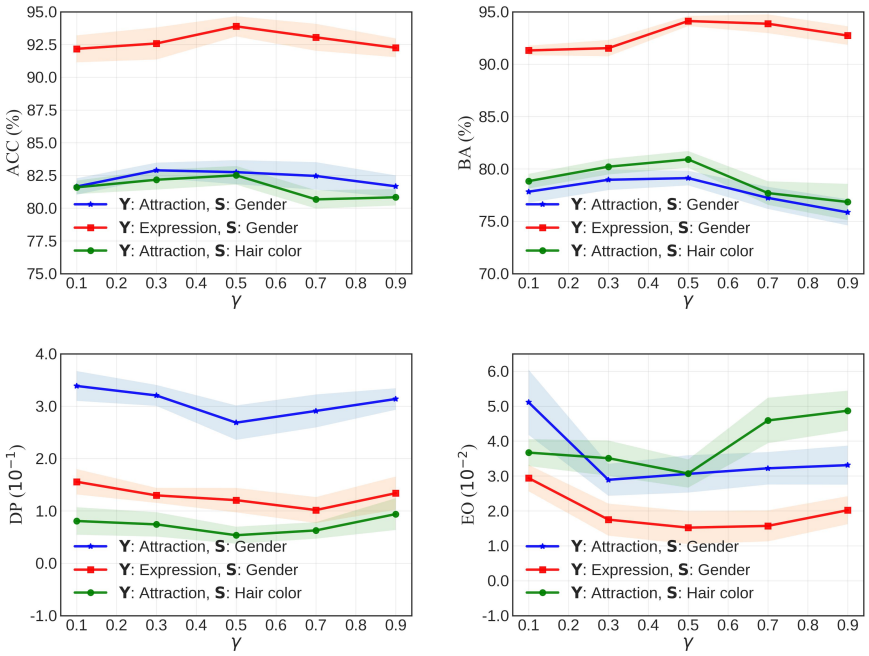

Fig. B3: Impact of γ. Shown is the mean $\pm$ standard deviation of 3 runs with different random seeds.

3. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)