WUHAN UNIVERSITY

UCIRVINE

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

Keywords: Vision Transformer · Accuracy · Fairness · Adaptive Masking

ECCV 2024

Introduction

Fairness in ViT is investigated in several recent works, yet a majority of them either sacrifice accuracy for fairness, or require a huge amount of computational cost. TADeT [1], a targeted alignment technique, seeks to identify and eliminate bias from the query matrix in ViT, but this method sacrifice accuracy for fairness. Debiased Self-Attention (DSA) [2] is a fairness-target approach that enforces ViT to eliminate spurious features correlated with the sensitive label. However, it requires costly two-stage training, which is hard to deploy in real scenarios.

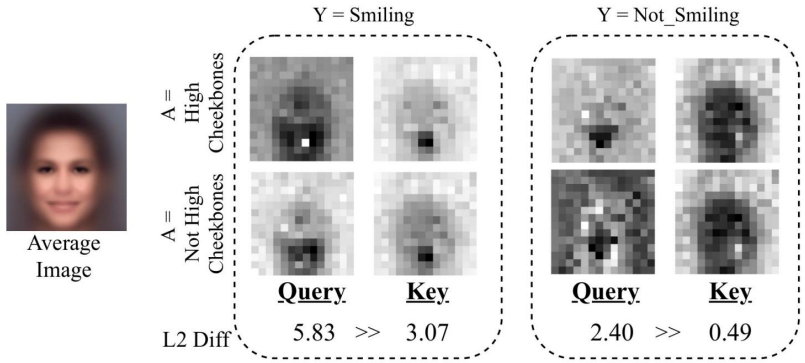

Fig a. The effect of TADeT.

Fig b. The effect of DSA.

To address the aforementioned challenges, we propose FairViT aiming at addressing fairness and accuracy concerns. These are our contributions:

• We introduce an adaptive masking framework wherein group-specific masks. and weights are learned to enhance fairness. We equip the adaptive masking with a backward algorithm that optimizes the masks and weights.

• We incorporate an extendable distance loss function manipulating the output scores to augment accuracy.

• We conduct extensive experiments on real datasets and demonstrate FairViT achieves accuracy better than alternatives, even with competitive computational efficiency. Furthermore, FairViT attains appreciable fairness results.

Overview

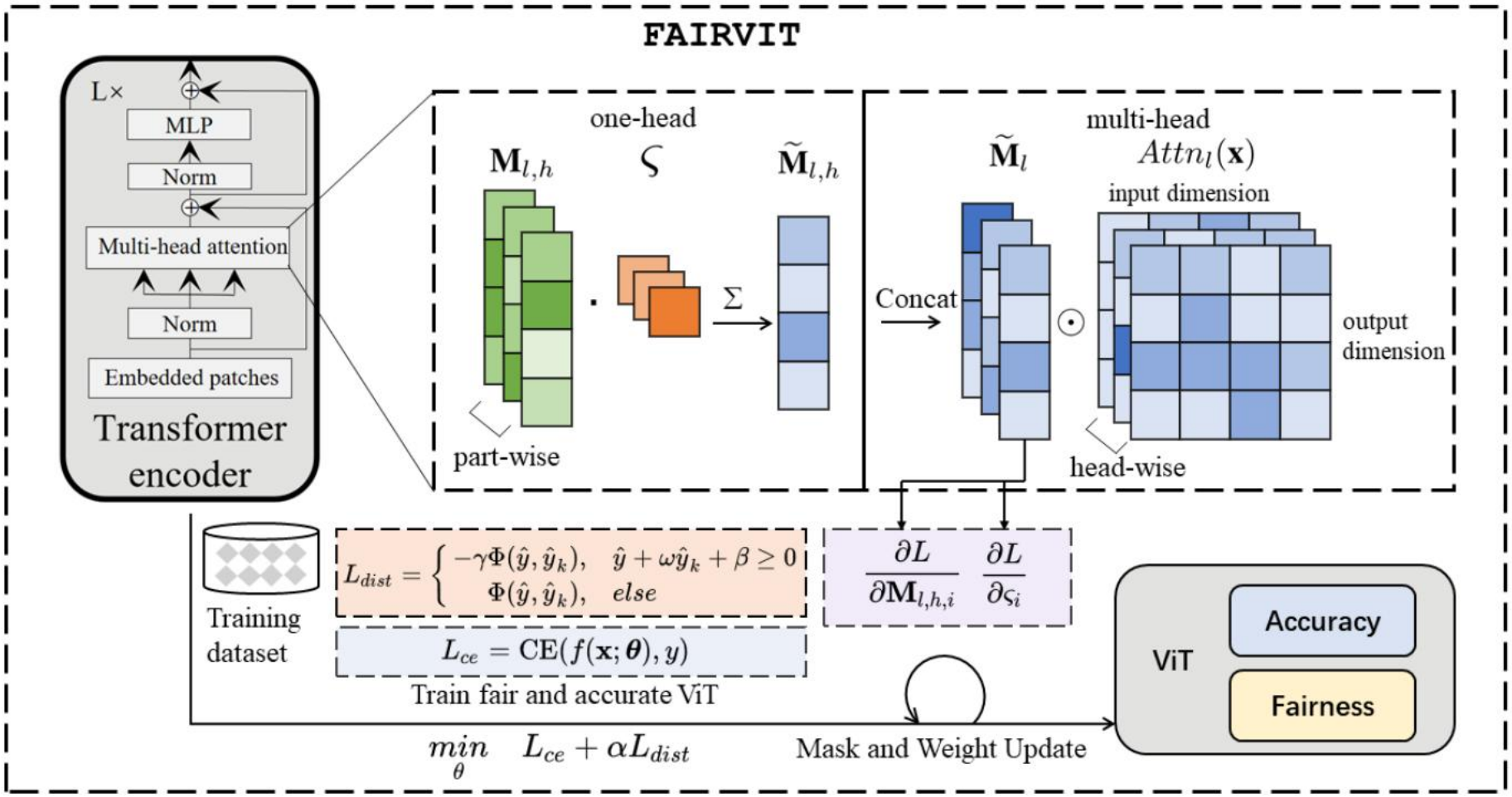

Fig. 1: An illustration of FairViT . For the forward propagation, we first apply weight ς to $\mathbf{M}_{l,h}$, calculate the weighted sum $\widetilde{\mathbf{M}}_{l,h}$, which is utilized to assist attention mechanism to control the information flow. For the backward propagation, we optimize $\mathbf{M}_{l,h,i}$ and ς_i. Additionally, we introduce a novel distance loss L_{dist}.

Adaptive masking

We associate each part (splitted by sensitive groups of the dataset) with a corresponding mask and weight. Each part i has a corresponding mask $\mathbf{M}_{l,h,i}$ and weight ς_i as parameters.

$$\mathrm{Attn}_{l,h}(\mathbf{x}) = \mathrm{S}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{3}$$

$$\widetilde{\mathbf{M}}_{l,h} = \sum_{i=1}^{G}(\varsigma_i \mathbf{M}_{l,h,i}) \tag{4}$$

$$\mathrm{HA}(\mathbf{x}, \mathbf{M}_{l,h}) = \widetilde{\mathbf{M}}_{l,h} \odot \mathrm{Attn}_{l,h}(\mathbf{x}) \tag{5}$$

where $\odot$ is the element-wise product, $\mathbf{M}_{l,h,i}$ represents the i_{th} mask ($i \in \{1, \ldots, G\}$) within layer l and head h, ς_i is the weight of $\mathbf{M}_{l,h,i}$, and $\widetilde{\mathbf{M}}_{l,h}$ is the weighted sum of $\mathbf{M}_{l,h,i}$.

Backward optimization

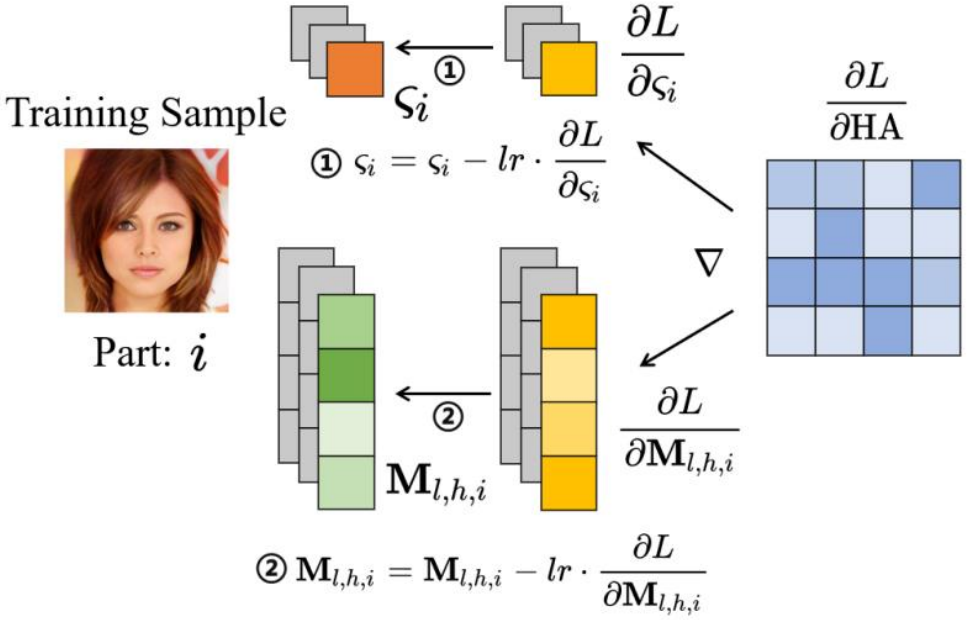

Fig. 3: An illustration of the update process. We ascertain the specific part i to which the training sample belongs, and ∇ refers to the gradient calculation, specified in (7) and (8). The gray blocks signify that the gradients are zero during the backward pass of this training sample.

$$\frac{\partial L}{\partial \mathbf{M}_{l,h,i}} = \begin{cases} \frac{\partial L}{\partial \mathrm{HA}} \mathrm{Attn}_{l,h}(\mathbf{x}) \cdot \varsigma_i, & \text{if} \quad i = g \\ \mathbf{0}, & \text{otherwise,} \end{cases} \tag{7}$$

$$\frac{\partial L}{\partial \varsigma_i} = \begin{cases} \sum_p \sum_d \left(\frac{\partial L}{\partial \mathrm{HA}} \mathrm{Attn}_{l,h}(\mathbf{x}) \cdot \left(\sum_d \mathbf{M}_{l,h,i} \right) \right), & \text{if} \quad i = g \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

Baselines

Table 1: The performance of image classification on CelebA dataset [15] with Vanilla [7], TADeT-MMD [25], TADeT [25], FCSL [19], FSCL+ [19] and FairViT. Shown is the mean of 3 runs with different random seeds. Highlighted is the best result.

method	**Y**: Attraction, **S**: Gender				**Y**: Expression, **S**: Gender				**Y**: Attraction, **S**: Hair color			
	$ACC_{\%}$	$BA_{\%}$	EO_{e-2}	DP_{e-1}	$ACC_{\%}$	$BA_{\%}$	EO_{e-2}	DP_{e-1}	$ACC_{\%}$	$BA_{\%}$	EO_{e-2}	DP_{e-1}
Vanilla	74.01	72.36	14.43	3.245	88.42	88.85	4.91	1.489	76.48	74.55	3.61	1.896
TADeT-MMD	79.89	73.85	7.10	3.693	92.51	93.03	2.48	1.290	77.97	75.64	2.27	1.491
TADeT	78.73	74.52	3.11	3.116	90.05	90.68	4.86	1.443	78.49	77.42	3.78	1.057
FSCL	79.09	74.76	1.78	3.004	89.37	90.08	1.76	1.344	78.85	78.06	2.65	0.989
FSCL+	77.26	73.42	**0.79**	**2.604**	88.83	89.02	**1.20**	1.263	78.02	77.37	**1.79**	0.834
FairViT	**84.01**	**79.96**	1.15	2.837	**94.27**	**94.12**	1.52	**1.205**	**82.52**	**81.56**	2.10	**0.701**

In comparison to FSCL+ [3], which is our main competitor, FairViT achieves a significantly higher accuracy of at least 4.5%. In terms of fairness metrics, FairViT showcases excellent unbiased effects.

Interpretability Study

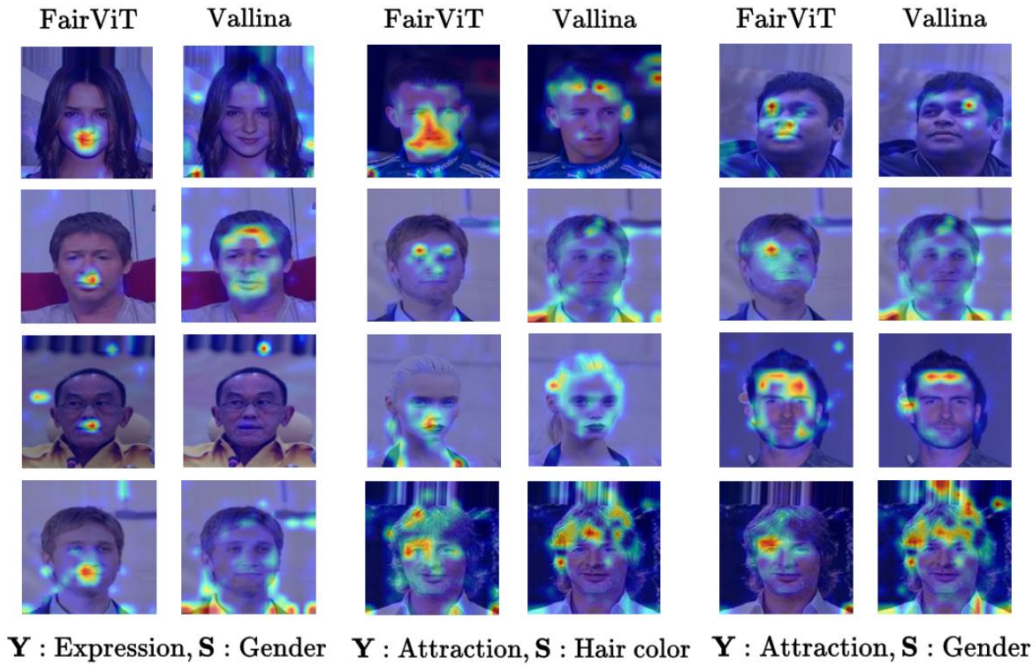

Fig. 5: The interpretability study of FairViT .

The Vanilla method appears to capture information relevant to sensitive attributes. In contrast, FairViT, leveraging adaptive masking, demonstrates a tendency to extract information relevant to the target attributes. Furthermore, FairViT generates heat maps [4] that are distributed more distinctly and densely in space, potentially indicating enhanced model learning.

Thank you!

Presenter: Bowei Tian
Personal website: https://bowei.netlify.app
OCRID: Bowei Tian (0009-0005-7275-7955) - ORCID

References

[1] Sudhakar, S., Prabhu, V., Krishnakumar, A., Hoffman, J.: Mitigating bias in visual transformers via targeted alignment. arXiv preprint arXiv:2302.04358 (2023).

[2] Qiang, Y., Li, C., Khanduri, P., Zhu, D.: Fairness-aware vision transformer via debiased self-attention. arXiv preprint arXiv:2301.13803 (2023).

[3] Park, S., Lee, J., Lee, P., Hwang, S., Kim, D., Byun, H.: Fair contrastive learning for facial attribute classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10389–10398 (2022)

[4] Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. arXiv preprint arXiv:2005.00928 (2020)