Supplementary Material: Mitigating Bias in Visual Transformers via Targeted Alignment

Sruthi Sudhakar
sruthis@gatech.edu
Viraj Prabhu
virajp@gatech.edu
Arvindkumar Krishnakumar
akrishna@gatech.edu
Judy Hoffman
judy@gatech.edu

Georgia Institute of Technology
Atlanta, GA

Contents

1 Evaluation Setting Details

In Sec 4.1 we mention that we choose three settings to benchmark prior work and evaluate TADeT. These three settings, described as a tuple (y,a), are as follows: i) (Smiling, High Cheekbones), ii) (Wavy Hair, Male), and iii) (Wavy Hair, Wearing Lipstick). Here we provide details on the choice of these three settings.

1.1 Smiling with High Cheekbones

The first setting we evaluate on is (Smiling, High Cheekbones) (See Fig 1). We notice that there is a significant skew in the distribution of the CelebA dataset, where most "Smiling" faces are correlated with "High Cheekbones" and most "Not Smiling" faces are correlated with "Not High Cheekbones". However, we understand that not all such correlations will necessarily translate to a *model* bias. Therefore, we analyze the true positive rate and false positive rate of a transformer trained on this dataset to predict the "Smiling" attribute, for each setting of the protected attribute "High Cheekbones". We obtain the following results: $\text{TPR}_{a=1} = 92.83\%$, $\text{TPR}_{a=0} = 67.76\%$, $\text{FPR}_{a=1} = 9.65\%$, and $\text{FPR}_{a=0} = 2.94\%$. Clearly, the model performs significantly worse on correctly predicting "Smiling" when the individual does *not* have "High Cheekbones" ($a = 0$). Furthermore, the model has a 6.71% larger FPR for the "High Cheekbones" group ($a = 1$). This confirms that the model indeed has a strong bias of (spuriously) correlating the presence of "High Cheekbones" with whether they are "Smiling". Due to this clear bias, we choose (Smiling, High Cheekbones) as our first setting for evaluation.

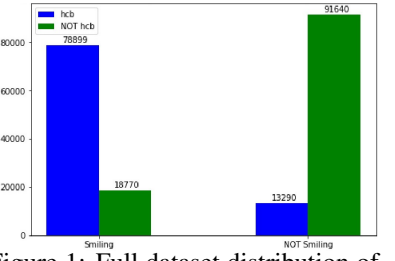

Figure 1: Full dataset distribution of task, protected attribute tuple (Smiling, High Cheekbones).

1.2 Wavy Hair with Male

Next, we evaluate on (Wavy Hair, Male) (See Fig 2). First, we notice that the dataset contains less "Male" individuals (28.5%) than "Not Male" individuals (71.5%). Due to the fact that "Male" is underrepresented in the data, we suspect that the model will not be able to learn as strong of a representation of the images with the "Male" attribute as it will for the "Not Male" attribute, which could lead to some type of bias. Next, we notice that most "Male" individuals are labeled as "Not Wavy Hair", which could cause a spurious correlation between these two attributes. When analyzing a Transformer trained on the "Wavy Hair" prediction task, we notice exactly that: a high difference in TPR where $\text{TPR}_{a=1} = 36.06\%$ and $\text{TPR}_{a=0} = 67.92\%$. This indicates that a (spurious) dataset correlation of "Male" with "Not Wavy Hair" is being learned by the model, leading to a large bias wherein the model is overpredicting "Male" to have "Not Wavy Hair".

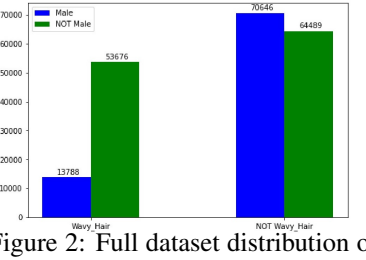

Figure 2: Full dataset distribution of task, protected attribute tuple (Wavy Hair, Male).

1.3 Wavy Hair with Wearing Lipstick

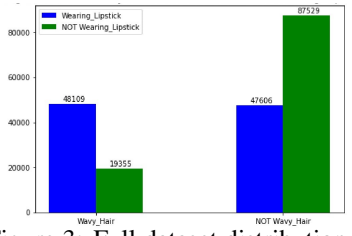

Figure 3: Full dataset distribution of task, protected attribute tuple (Wavy Hair, Wearing Lipstick).

Finally, we evaluate on (Wavy Hair, Wearing Lipstick) (See Fig 3). There is a natural correlation in the dataset where most "Male" individuals are "Not Wearing Lipstick", and vice versa. Since "Wearing Lipstick" and "Male" are correlated, and "Male" and "Wavy Hair" are correlated (as shown in 1.2), we want to see if the bias that is present when the protected attribute is "Male" will persist when we set the protected attribute as "Wearing Lipstick". Looking at the dataset distributions themselves, we notice that the number of people "Wearing Lipstick" and "Not Wearing Lipstick" is closer to 50/50 than in the previous setting (protected attribute="Male"). Next, we notice that there is a correlation in the dataset of people with "Wavy Hair" and people "Wearing Lipstick". After training a transformer, we notice that the Equalized Odds and Balanced Accuracy Difference are almost as high as the (Wavy Hair, Male) setting. Therefore, we conclude that this setting will be a good test-bed, as the distribution is not as clearly skewed as (Wavy Hair, Male), but there is poor performance on both fairness metrics Equalized Odds and Balanced Accuracy Difference.

2 Balanced Accuracy Difference Use Cases

In Sec. 4.1 of the main paper, we introduce Balanced Accuracy Difference, a fairness metric that shares some of the motivation behind accuracy equity [1], but with an important implementation difference to account for real-world data distributions. While accuracy equity suggests taking the difference in Standard Accuracy across the protected attribute, we take the difference in Balanced Accuracy (our performance metric) across the protected attribute. By doing so, we can account for class imbalance in the dataset, as we saw in Sec. 1. Furthermore Balanced Accuracy Difference is important because it provides a more holistic understanding of the Equalized Odds metric, as we now elaborate.

Consider a situation wherein after debiasing a model, the true positive rate (TPR) difference across a protected attribute *increases* slightly while false positive rate (FPR) difference decreases substantially. Since Equalized Odds is an average of TPR difference and FPR difference (Equalized Odds $= \frac{1}{2}[TPR_{a=1} - TPR_{a=0}] + \frac{1}{2}[FPR_{a=1} - FPR_{a=0}]$), the resulting Equalized Odds measure will reduce, indicating that the model is fairer than the original model. However due to the *increased* TPR difference, predicting the positive outcome for the protected attribute is actually *more* unfair than before debiasing! This means that for the positive outcome, a large bias across the protected attribute still exists, which the Equalized Odds metric does not adequately capture.

However, in this situation, the Balanced Accuracy Difference will be high as it looks at the *differences* in Balanced Accuracy within each subgroup. More specifically, Balanced Accuracy Difference can be rewritten as:

$$\text{Balanced Acc. Difference}(\Delta\text{BA}) = \frac{1}{2}[TPR_{a=0} + TNR_{a=0}] - \frac{1}{2}[TPR_{a=1} + TNR_{a=1}] \quad (1)$$

$$= \frac{1}{2}[TPR_{a=1} + TNR_{a=1}] - \frac{1}{2}[TPR_{a=0} + TNR_{a=0}] \tag{2}$$

$$= \frac{1}{2}[TPR_{a=1} - TPR_{a=0}] + \frac{1}{2}[(TNR_{a=1} - TNR_{a=0}] \tag{3}$$

$$= \frac{1}{2}[TPR_{a=1} - TPR_{a=0}] + \frac{1}{2}[(1 - FPR_{a=1}) - (1 - FPR_{a=0})] \tag{4}$$

$$= \frac{1}{2}[TPR_{a=1} - TPR_{a=0}] + \frac{1}{2}[(FPR_{a=0} - FPR_{a=1})] \tag{5}$$

$$= \frac{1}{2}[TPR_{a=1} - TPR_{a=0}] - \frac{1}{2}[(FPR_{a=1} - FPR_{a=0})] \tag{6}$$

By Eq. 6, it is clear that given our situation where TPR difference is slightly higher than before debiasing, and FPR difference is lower than before debiasing, Balanced Accuracy Difference would *increase*, indicating that the model is behaving in a biased manner, even though the Equalized Odds measure decreases. Therefore, a user will realize that the drop in Equalized Odds does not tell the full story, as Balanced Accuracy Difference will indicate that their model still encodes a bias, especially towards predicting the positive value for one setting of the protected attribute. Hence, we advocate for using Balanced Accuracy Difference as an *additional* fairness metric, along with Equalized Odds.

3 Transformer Feature Visualizations

Recall that in Figure 1 of the paper, we presented a visualization of the average Query and Key activations for each (y,a) tuple combination, for a specific attention head and channel of a transformer trained for the "Smiling" prediction task. In Figure 4, we provide visualizations that demonstrate that the differences we notice in the query activations, and the similarity noticed in the key activations, generalizes across different attention heads and channels of the query and key matrices. Further, this also generalizes across different tasks.

4 Analyzing Class-Specific Alignment for CNNs

In `TADeT`, we propose using class-specific alignment wherein we align the distribution of the protected attribute within the task attribute by utilizing an adversary head *per-attribute* during adversarial learning. We have shown the benefits of this method for debiasing visual transformers. In Table 1, we show that such class-specific alignment for adversarial training improves upon previous debiasing algorithms in most settings for CNN's as well.

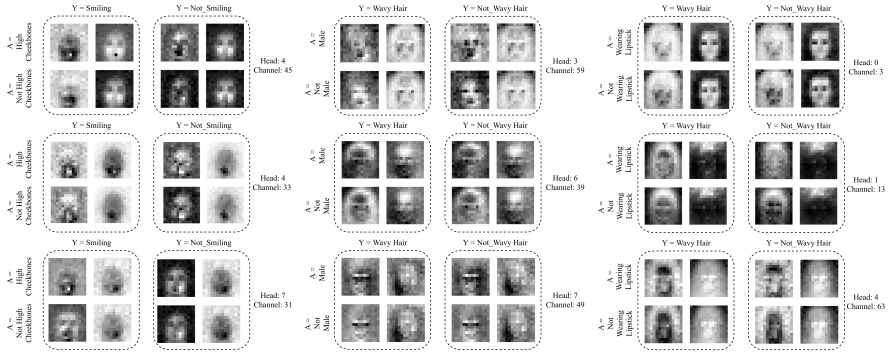

Figure 4: We show that the variance in the query matrix activations for a particular task label, y, across a protected attribute a, generalizes across attention heads, channels, and data settings. The Transformer Query/Key Matrix has 64 channels and 8 heads, and we have chosen 3 random channel/head combinations from all 3 tasks to depict the generalization of differences in activations.

	Y: Wavy Hair **A:** Male				**Y:** Smiling **A:** High Cheekbones			
Method	EO ↓	Δ BA (%) ↓	BA (%)↑	Acc (%)↑	EO ↓	Δ BA (%) ↓	BA (%)↑	Acc (%)↑
Original CNN	16.71	8.08	77.99	82.20	14.66	2.69	88.15	93.06
DANN [2]	14.75	7.39	77.36	81.12	15.04	**1.85**	87.97	93.03
DANN Class Specific	**14.57**	**7.06**	77.21	80.89	**14.51**	3.37	88.40	93.24

Table 1: CNN Debiasing results. Y=Task. A=Protected Attribute. EO=Equalized Odds. ΔBA=Balanced Accuracy Difference. BA=Balanced Accuracy.

References

[1] William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity, 2016.

[2] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation, 2015.