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

Abstract

As transformer architectures become increasingly prevalent in computer vision, it is critical to understand their fairness implications. We perform the first study of the fairness of transformers applied to computer vision and benchmark several bias mitigation approaches from prior work. We visualize the feature space of the transformer self-attention modules and discover that a significant portion of the bias is encoded in the query matrix. With this knowledge, we propose `TADeT`, a targeted alignment strategy for debiasing transformers that aims to discover and remove bias primarily from query matrix features. We measure performance using Balanced Accuracy and Standard Accuracy, and fairness using Equalized Odds and Balanced Accuracy Difference. `TADeT` consistently leads to improved fairness over prior work on multiple attribute prediction tasks on the CelebA dataset, without compromising performance.

1 Introduction

Over the past decade, deep learning-based computer vision has been advancing at a rapid pace and is being deployed in high-stakes applications ranging from candidate job hiring to facial recognition systems. Several recent studies have shown that deep models take advantage of spurious correlations between attributes in the data when learning to make predictions. Such unintended correlations can cause the model to act in a biased way, such as having lower accuracy on certain sub-populations of the data [4, 7, 12, 20, 27]. These biases can be harmful and cause discrimination towards these populations once deployed.

While the fairness of computer vision tasks learnt using convolutional neural network (CNN) models is well explored, the fairness implications of *Visual Transformers* are yet to be studied. Transformers were first introduced in NLP [22], and later found to be effective on computer vision tasks [6] due to their ability to learn long-range relationships, unlike CNN's which have have a limited field of view at any given layer. There has been a rise in the use of transformers in vision tasks [13]. In this paper, we perform the first study of visual transformers with respect to fairness. Transformers have a naturally separated and spatially cohesive feature space [22], and we investigate how this unique architecture may be leveraged to ensure fairness.

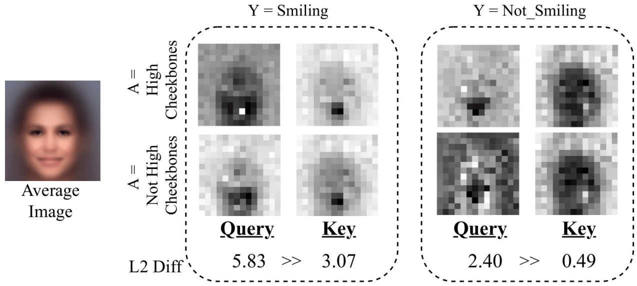

Figure 1: We discover that bias in a visual transformer model across a particular attribute label (e.g. A = high cheekbones / not high cheekbones) can be attributed to a variance in the query matrix activations for a fixed task label (Y = smiling / not smiling), even with similar key activations. This observation motivates `TADeT` (example shown on CelebA [15]).

There are several ways to measure the fairness of a model [1, 8, 11, 29]. In this work, we focus on an image classification task for which all predictive outputs are equally desirable, and hence measure performance using *balanced* (multiclass) accuracy [18]. Further, we focus on two fairness metrics: Equalized Odds [8], as it focuses on equal true positive and false positive rates across the protected attribute, and a new auxiliary measure we introduce, **Balanced Accuracy Difference** (Δ BA) which measures the difference in Balanced Accuracy across the two groups. This metric, which is similar to accuracy equity [1], gives practitioners an overall idea of the *difference* in model performance across a protected attribute.

Several debiasing algorithms have been proposed and studied for CNNs, including reweighting instances [24], MMD alignment [16], domain independent training [8, 25], and adversarial training [28]. However, many of these techniques suffer from a trade-off where increased bias mitigation results in reduced predictive performance [25]. In addition, these works either require a large number of the protected class data or are designed for learning with a CNN model, making it unclear how well these methods will translate to newer backbone architectures like visual transformers.

We perform the first fairness assessment of visual transformer models and benchmark existing bias mitigation algorithms. We then introduce **`TADeT`**, Targeted Alignment for Debiasing Transformers, a novel algorithm that applies class-specific Maximum Mean Discrepancy (MMD) [19] alignment and class-specific adversarial [28] debiasing, to a *targeted* part of the visual transformer feature space. We discover that when bias exists for a particular attribute, the core piece of the model which differs between two subgroups (one with and one without said attribute) can be observed through the query activations. For example, in Figure 1 we consider a Transformer trained to predict the presence of the "Smiling" attribute, and visualize its *average* query and key matrix activations over the dataset. We find that for the subset of images where smiling is present, a large discrepancy exists in average query activations across the protected attribute of "High cheekbones" (box 1, column 1), whereas the key activations are much more similar (box 1, column 2). This result is also observed in the case of *not* smiling images (box 2). Thus, by imposing alignment [10] directly on the query activations we may be able to effectively reduce model bias while limiting the impact on predictive performance.

We evaluate our method on the task of single-label attribute prediction on the CelebA dataset [15], while requiring invariance to a defined protected attribute. We measure the performance of our method using Balanced Accuracy and Standard Accuracy, and the fairness of our method using Equalized Odds [11] and Balanced Accuracy Difference. We make the following contributions:

1. We perform the first assessment of the fairness of representations learned by visual transformers, and benchmark several existing debiasing methods from the literature.
2. We visualize the transformer feature space and discover that a significant portion of learned bias is encoded in the activations of the transformer's query matrices.
3. Motivated by this, we propose `TADeT`, a debiasing strategy for visual transformers using *targeted*, *class-specific*, MMD alignment and adversarial debiasing to improve upon existing fairness algorithms without compromising accuracy.

2 Related Work

Transformers in computer vision. Transformers were first introduced in NLP by Vaswani *et al.* [22], which used self-attention to develop better context and allow models to attend to relevant words across the entire corpus of text. Recently, ViT (Visual Transformer) [6] has emerged as a new architecture for vision capable of achieving state-of-the-art performance on standard recognition tasks. Given the strong performance of this architectural paradigm, many future works are bound to leverage the visual transformer model, necessitating a study into the fairness implications and mitigation approaches designed for transformers.

Measuring Biases. Several metrics exist for measuring bias in machine learning models. Demographic parity [8] measures whether the rates of positive outcomes across all protected groups are equal, making it susceptible to the self-fulfilling prophecy and subset targeting. Bias Amplification [29] measures how a model *amplifies* biases, though it does not provide feedback about the inherent biases common in visual recognition datasets [2, 5, 21, 24]. Equalized Odds [11], one of the metrics we report, offers a stricter definition of Equality of Opportunity, by adding an extra constraint on the Demographic Parity measure where each group must receive the positive and negative outcomes at equal rates if the datapoints in this group qualify for that outcome. Finally accuracy equity [1] compares accuracy across subpopulations. However, when a dataset is class imbalanced, this measurement will have an implicit bias against the minority class. Therefore, we introduce Balanced Accuracy Difference as a way to measure the difference in a model's performance across a protected attribute while accounting for class imbalance.

Mitigation Methods. Some of the most straightforward mitigation methods focus on re-sampling or re-weighting techniques [3, 9], but these methods may have limited information gain through resampling and in some cases suffer from overfitting [26]. A different style of approach seeks to directly impose alignment of feature spaces by optimizing an MMD loss between features from different subgroups [16], though mitigation benefits are variable [17]. A domain discriminative [8] approach trains a classifier for each protected attribute and task combination and hence attempts to induce 'fairness through awareness', though recent works [25] demonstrate limitations including a lack of leveraging between-class information and an increased number of parameters. A domain independent approach [24] trains an ensemble of classifiers with a shared feature representation to ensure that rare attribute values are adequately represented. While this approach can lead to large bias reductions, it is limited in that it assumes adequate labeled data for each setting of the protected attribute in order to train a representative model.

Many recent approaches have leveraged the success of adversarial learning techniques for domain alignment [10]. Adversarial Debiasing [28] alternates between training a classifier to distinguish the protected attribute and updating the representation to be invariant to the protected attribute. LAFTR [17] is an adversarial method with the ability to modify the training objective based on the desired fairness measure (equalized odds, equality of opportunity, and demographic parity). However, in our experiments using this method with the equalized odds objective, we find that it is unable to consistently achieve fairness improvements while maintaining high performance.

We benchmark MMD, Domain Independent, Adversarial Debiasing, and LAFTR methods on transformers. We then propose `TADeT`, a targeted alignment approach to debias transformers (leveraging the query matrix features) that achieves improved fairness with minimal reduction in predictive performance.

3 Approach

Recently, visual transformers [6] have emerged as a popular architecture for learning visual recognition models. While the fairness of CNN models has been documented [4] and mitigation strategies have been proposed [17, 28], we study the bias which is present within a visual transformer-based model and propose `TADeT`, a new bias mitigation algorithm designed for debiasing visual transformers. `TADeT` focuses alignment on query activations within the transformer feature space to remove bias against a protected attribute.

3.1 Notation and Background

Let $\boldsymbol{x}$ denote an input image and y the corresponding task label. Our goal is to learn a classifier h, on top of deep encoder, f, such that $h(f(\boldsymbol{x})) \rightarrow y$ achieves high heldout task accuracy while being fair with respect to a protected attribute, a. Fairness has many definitions; in this work we focus on designing a model which is invariant with respect to the protected attribute (see Sec 2).

The model f employs the visual transformer architecture introduced in Dosovitskiy *et al.* [6]. The input image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times C}$ is reshaped into N patches $\{x_1, .., x_N\}$, where each patch x_i is of size $\frac{HW}{N} \times \frac{HW}{N} \times C$. Each patch is then flattened and mapped to D dimensions with a trainable linear projection. A final learnable class token x_{class} of the same dimension, D, is appended to the sequence of patches. This class token is randomly initialized and updated throughout training. These $N+1$ patch embeddings are appended with a positional embedding to retain positional information, and then fed through L layers of the transformer encoder. Each layer of the encoder is identical and includes a multi-headed self-attention module followed by alternating MLP blocks and layernorms.

The multi-headed self-attention module includes M attention heads. For each attention head, m, the concatenated $N+1$ patch embeddings are split into 3 branches and each fed to a linear layer. The 3 branches result in learning Query (Q), Key (K), and Value (V) activations respectively, each of dimension $M \times (N+1) \times D$, which attempt to encode specific pieces of information. The value matrix is tasked with learning an encoding of the actual attributes of the data (*eg.* for a face image, information such as hair color, eye color, etc.). The key matrix encodes the location of *where* specific information is stored in the value matrix (*eg.* where the eye color is located). Finally, the query matrix is trained to "ask" the key matrix for the information it requires for the task. Activations produced from these 3 matrices are then combined to produce the self-attention defined in Vaswani *et al.*[22] as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{D}}\right)V \tag{1}$$

The activations from the self-attention module are fed into alternating MLP and layernorms in each layer of the encoder. Finally, the features for the class token, x_{class}, are extracted from the output of the final layer transformer encoder and used as the feature representation of input $\boldsymbol{x}$ to predict output y.

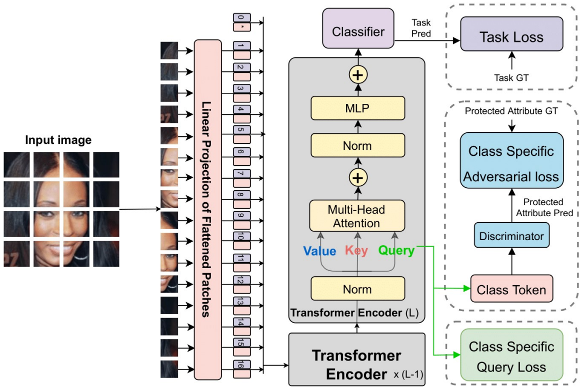

Figure 2: Overview of `TADeT`. First, a preprocessed input image and an additional class token are passed through L transformer encoder layers, and features extracted from the class token are used for task classification. Second, query activations of the last layer of the encoder are used for Targeted Debiasing in two ways: i) query activations corresponding to the class token are used for class-specific adversarial alignment (Sec. 3.2.1), and ii) for each class, the L_2 distance between average query activations for each setting of the protected attribute is minimized (Sec. 3.2.2).

3.2 `TADeT`: Targeted Alignment for Debiasing Transformers

Unlike CNNs, transformers possess a naturally interpretable feature space due to separation into queries and keys. On studying the average query and key activations of the last layer of a visual transformer encoder before making a prediction, we find a measurable difference across average query activations across the protected attribute values, but only minimal differences between average key activations across the protected attribute values (see Figure 1). These differences generalize across the D dimensions and M attention heads of the query matrix. This observation of a large difference in query activations motivates our focus on alignment of query activations.

We introduce `TADeT`, Targeted Alignment for Debiasing Transformers, a bias mitigation algorithm that focuses on aligning *query activations* (see Figure 2). `TADeT` considers a visual transformer model trained using a standard task loss and performs debiasing through the introduction of two class-specific alignment losses, 1) an adversarial loss which aligns the class token of the query activations, L_{adv}, (Sec 3.2.1) and, 2) an L_2 loss which penalizes large deviations between the average query activations for values of the protected attribute, L_q, (Sec 3.2.2). Our overall model considers the joint optimization between these two objectives together with a task supervised cross-entropy loss, L_{CE}.

3.2.1 Class-specific Adversarial Alignment

We begin alignment across populations with different attributes by leveraging adversarial mitigation [28]. In adversarial training for feature alignment, an adversary g is applied to a feature space to determine if two groups are distinguishable. Optimization alternates between training the discriminator to distinguish the groups, a in our case, and updating the feature encoder, f, to confuse the discriminator. Whereas prior work aligns the penultimate features for an instance, $\boldsymbol{x}$, we seek to align the class token of the query matrix, $Q_{\boldsymbol{x}}[:, N+1, :]$. This results in a standard min/max optimization over this targeted feature space:

$$\mathcal{L}_{\text{adv}}(\boldsymbol{x}; f, g) = \max_{f} \min_{g} \mathbb{E}_{\boldsymbol{x},y,a}[\mathcal{L}_{\text{CE}}(g(Q_{\boldsymbol{x}}[:, N+1, :]), a)] \tag{2}$$

Furthermore, prior works propose aligning the features across the protected attribute by aligning $p(\hat{y}=y|a=0)$ and $p(\hat{y}=y|a=1)$. However, this does not take into account the ground truth task label y, allowing two datapoints with different task labels to be aligned, which will lead to incorrect classification. Therefore, we propose using *class-specific* alignment where we align the features based on both the protected attribute value *and* the task label value (*i.e.* we only align $p(\hat{y}=y|a=0,y=1)$ with $p(\hat{y}=y|a=1,y=1)$, and only align $p(\hat{y}=y|a=0,y=0)$ with $p(\hat{y}=y|a=1,y=0)$).

3.2.2 Direct Alignment of Class-Specific Query Activations

Our hypothesis is that by reducing the differences between the average query activations, we can encourage invariance to the protected attribute and improve fairness. Following this we add a Maximum Mean Discrepancy (MMD) loss [19] to our objective function that minimizes the L_2 distance between the average query activations for instances from each attribute. We exclude the class token of the query activation, $Q_{\boldsymbol{x}}[:,N+1,:]$, as it does not encode spatial information. We first compute the average query activation in each minibatch, B, for each task label (y) and protected attribute (a) combination, where $Q_{\boldsymbol{x}_i}$ indicates the query activation for image $\boldsymbol{x}_i$:

$$\hat{Q}_{y,a} = \frac{1}{|B|} \sum_{i \in B, \text{ s.t. } y_i=y, a_i=a} Q_{\boldsymbol{x}_i}[:,1:N,:] \tag{3}$$

Next, we compute the average of L_2 differences across the protected attribute for each element in $\hat{Q}_{y,a}$ independently, and then average these differences across the D channels and M attention heads of the matrix. Due to the reasoning provided in Sec.3.2.1, we utilize class-specific alignment of the query loss to ensure that the appropriate features are aligned. For binary attributes this yields the following alignment loss:

$$\mathcal{L}_q(\boldsymbol{x};f) = \frac{1}{2 \cdot MD} \sum_{m \in M} \sum_{d \in D} \sum_{y \in \{0,1\}} ||\hat{Q}_{y,0}[m,:,d] - \hat{Q}_{y,1}[m,:,d]||_2 \tag{4}$$

Altogether, our model results from the following joint minimization:

$$\mathcal{L}(\boldsymbol{x},y;f,h,g) = \mathcal{L}_{\text{CE}}(h(f(\boldsymbol{x})),y) + \alpha\mathcal{L}_{\text{adv}}(\boldsymbol{x};g,f) + \beta\mathcal{L}_q(\boldsymbol{x};f) \tag{5}$$

where hyper-parameters α and β are tuned and optimized for fairness.

4 Experiments

4.1 Setup & Implementation Details

We evaluate `TADeT` on CelebA [15], a dataset containing 200k celebrity face images with annotations for 40 binary attributes. We present results on three settings, each with a corresponding binary task (y) that the model is trained to predict, and a binary protected attribute (a) over which we wish the model to be unbiased. The 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). We take careful steps to choose these settings: First we train a baseline Transformer model on different task attributes (including Smiling and Wavy Hair). Then we measure the widely-used Equalized Odds fairness metric to estimate how much bias is correlated with each protected attribute (results for Smiling reported in Figure 3). Finally, we choose the protected attribute with the highest Equalized Odds and thus the most bias (more details in the supplement).

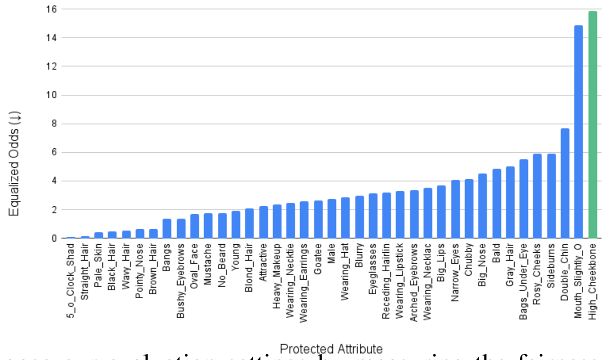

Figure 3: We choose our evaluation settings by measuring the fairness of a Transformer model trained to predict a given task attribute across different protected attributes. This figure above shows the fairness (measured by Equalized Odds) across each protected attribute for a Transformer trained to predict Smiling. We select High Cheekbones (in green) as it is the attribute with the greatest bias. A similar process is followed for the (Wavy Hair, Male) and (Wavy Hair, Wearing Lipstick) settings.

Implementation Details. We first train a transformer from scratch on each of these 3 settings. The transformer consists of 196 patches (each representing a 16x16 area of the image), 1 class token patch, 12 transformer encoder layers, and 8 attention heads. We flatten and project each patch into a 64-dimensional vector and add positional embeddings [25]. The embedded patches are fed into the transformer encoder. We use parameter sharing across key and value weights [23]. After the transformer encoder processes the patch embeddings, the class token patch is fed into 2 fully-connected layers and a sigmoid layer to produce a single normalized output score (since we deal with binary classification). We use the Adam optimizer [14] with a learning rate of 1×10^{-4}, train for 50 epochs, and following Wang *et al.*[25], select the model with the best weighted average precision (AP) on the validation set.

We implement `TADeT` by applying a class-specific adversarial and class-specific query loss on top of our pretrained transformer model to perform targeted debiasing. We align the query class token activations by applying one adversary head per setting of the task label. The adversary takes in the 512-dimensional vector query class token of the final (12^{th}) layer of the transformer encoder. The class token is processed through 2 fully-connected layers and a sigmoid layer to produce a single normalized output score of the binary protected attribute a. We perform a sweep across multiple loss weights for the adversarial (α) and query (β) losses, and select the one yielding the best fairness metrics.

Metrics. As performance metrics, we report Standard Accuracy and Balanced Accuracy [18], $\frac{1}{4}[TPR_{a=0}+TNR_{a=0}+TPR_{a=1}+TNR_{a=1}]$, which is needed when task data is imbalanced. As fairness metrics, we report Equalized Odds [11] and introduce Balanced Accuracy Difference. Equalized Odds is used to understand the discrepancy in true positive rates and false positive rates across the protected attribute. Balanced Accuracy Difference looks at the difference in Balanced Accuracy between the binary values of the protected group. By computing the difference in Balanced Accuracy, as opposed to difference in Standard Accuracy proposed by Dieterich *et al.*[1], we are able to account for the class imbalance in the original dataset. A detailed use case is described in the supplement.

$$\text{Balanced Acc. Difference}(\Delta\text{BA}) = \frac{1}{2}[TPR_{a=0}+TNR_{a=0}] - \frac{1}{2}[TPR_{a=1}+TNR_{a=1}] \quad (6)$$

Method	Y: Smiling A: High Cheekbones				Y: Wavy Hair A: Male				Y: Wavy Hair A: Wearing Lipstick			
	EO ↓	Δ BA (%) ↓	BA (%)↑	Acc (%)↑	EO ↓	Δ BA (%) ↓	BA (%)↑	Acc (%)↑	EO ↓	Δ BA (%) ↓	BA (%)↑	Acc (%)↑
Transformer	15.89	9.18	87.00	92.21	19.76	12.10	72.79	78.18	19.72	9.86	74.33	78.81
MMD [16]	14.83	3.26	87.65	92.61	20.86	12.26	72.90	78.51	17.88	7.82	76.00	**79.98**
DANN [10]	15.13	3.58	87.67	<u>92.72</u>	19.74	9.36	74.70	79.65	18.63	7.67	75.21	79.31
LAFTR [17]	16.13	4.95	87.42	**92.78**	19.03	9.24	74.75	79.55	18.17	**6.50**	76.11	79.55
Domain Ind. [25]	**13.12**	8.59	86.61	90.89	**14.99**	11.01	69.71	74.09	**13.29**	9.03	70.61	73.80
TADeT (Ours)	<u>14.77</u>	**2.42**	**87.73**	92.68	<u>18.58</u>	**8.63**	**75.10**	**79.73**	<u>17.31</u>	<u>7.08</u>	**76.11**	<u>79.91</u>

Table 1: Debiasing results. Y=Task Attribute. A=Protected Attribute. EO=Equalized Odds. ΔBA=Balanced Accuracy Difference. BA=Balanced Accuracy. Best and second best performing method bolded and underlined respectively.

4.2 Baselines

We evaluate the following debiasing algorithms with the pretrained transformer as baselines.
i) Maximum Mean Discrepancy (MMD) [16]: MMD computes a mean of penultimate layer feature activations for each setting of the protected attribute, and then minimizes their L_2 distance. We select a loss weight that yields the highest validation results.
ii) Domain Adversarial Neural Network (DANN) [10]: Following successful adversarial debiasing [11], we build on domain adversarial alignment [10] using an attribute adversary learned on top of the penultimate layer activations. The adversarial head consists of 2 linear layers that take in a 512-dimensional vector class token, followed by a sigmoid, and select the adversarial loss weight that yields the highest validation results.
iii) LAFTR [17]: We train a model with a modified adversarial objective which attempts to satisfy the Equalized Odds fairness measure. This objective is implemented by minimizing the average absolute difference on each task attribute-protected attribute (y,a) combination. This allows the model to directly optimize for the equalized odds measure.
iv) Domain Independent Training [25]: We learn a shared feature representation with an ensemble of classifiers (two), trained with the same baseline hyperparameters. Given the predictions, we perform inference by averaging the class decision boundaries.

4.3 Results

We report results in Table 1. We find that TADeT consistently improves both Equalized Odds and Balanced Accuracy Difference while maintaining Balanced Accuracy and Standard Accuracy compared to the original transformer model (best or second best in 11/12 settings). Within a test setting there may be other debiasing methods which produce stronger debiasing results according to a single fairness metric, such as LAFTR on the (Wavy Hair,Wearing Lipstick) setting, but we note that across settings and considering all metrics our method consistently performs comparably or better than prior work. We further note that while Domain Independent training provides the strongest Equalized Odds performance, this approach does not produce strong Balanced Accuracy Difference (fairness) or Balanced Accuracy and Standard Accuracy (performance), likely due to imbalanced task data for the underrepresented protected attribute. For example, in the (Wavy Hair, Male) setting, though we see a 3.59% drop in Equalized Odds compared to TADeT, this is accompanied by worse performance on all other metrics including: a 2.38% increase in Balanced Accuracy Difference, a 5.39% drop in Balanced Accuracy, and a 5.64% drop in Standard Accuracy.

4.4 Ablating TADeT

In Table 2 we ablate TADeT and observe that each component of our method contributes to improved fairness measures.

#	Adversarial Loss						Query Loss		Fairness Metrics ↓		Performance Metrics ↑	
	activations		on top of		alignment		alignment					
	penultimate	query	full	class token	full	class spe	full	class spe	Eq Odds	Δ Bal Acc	Bal Acc	Std Acc
1* []	(Original Transformer Model)								15.89	9.18	87.00	92.21
2* []	✓		✓		✓				15.13	3.58	87.67	92.72
3	✓		✓			✓			14.64	3.46	88.01	92.90
4		✓	✓		✓				15.35	3.54	87.71	92.84
5		✓	✓			✓			14.93	3.89	87.59	92.57
6		✓		✓	✓				16.17	2.24	87.24	92.67
7		✓		✓		✓			14.34	2.98	87.90	92.69
8		✓		✓		✓	✓		15.10	2.40	87.48	92.55
9 (`TADeT`)		✓		✓		✓		✓	14.77	2.42	87.73	92.68

Table 2: Ablating `TADeT`. This analysis portrays the importance of applying a class specific adversary to align the class token activations across the protected attribute, and performing MMD alignment on the query matrix activations across the protected attribute. * indicates a baseline method we are comparing against.

⇒ **Class specific adversary improves fairness (Row 3):** We first experiment with applying a class-specific adversarial loss to align distributions within a setting of the task attribute. This leads to improvements in both fairness metrics and both accuracy metrics.

⇒ **Positioning adversary head on full query activations is ineffective (Row 4 & 5)**: Next we move the adversary head on top of the full output of the final query matrix. We experiment with both a standard (Row 4) and class-specific (Row 5) adversarial loss. While this does not do better than DANN, we hypothesize that this is because the input to the adversary is the entire query matrix rather than just the class token, which is what is actually used for task classification.

⇒ **Only operating on the query class token greatly improves fairness(Row 6 & 7)**: To test our prior hypothesis, we allow the adversary to only operate on the class token of the query matrix. We experiment with both a standard (Row 6) and class specific (Row 7) adversarial loss and notice that the latter results in strong Equalized Odds measures.

⇒ **Adding query loss improves Balanced Accuracy Difference but sacrifices Equalized Odds (Row 8):** Next, following our intuition of equalizing the query activations, we add a query loss. We notice an improvement in Balanced Accuracy Difference but no significant improvement in Equalized Odds. We again hypothesize that class specificity is required to notice improvements.

⇒ **Adding class specificity to query loss improves Equalized Odds while maintaining Balanced Accuracy Difference (Row 9)**: Finally, a class-specific query loss combined with a class-specific adversarial loss on the query class token achieves the strongest and most consistent performance across all metrics, implying that each component of `TADeT` is helpful in reducing bias while maintaining strong predictive performance.

4.5 Analysis

As seen in Figure 1 (*left*), before applying `TADeT`, the average query activations across both settings of the protected attribute are very different (large L_2 distance of 5.82 and 2.4 for smiling and not smiling respectively), while the average activations from the key matrix are much closer (L_2 distance of 3.07 and 0.49). These qualitative visualization results and quantitative L_2 results are consistent across different attention heads and channels of the query and key activations (additional results provided in the supplement).

After performing debiasing using `TADeT`, we visualize the new values of the average query and key activations and calculate the same L_2 distances. As seen in Figure 4, the query activations look very similar and have a small L_2 difference (0.01 and 0.0) across settings of the protected attribute. These results, combined with our improved fairness results

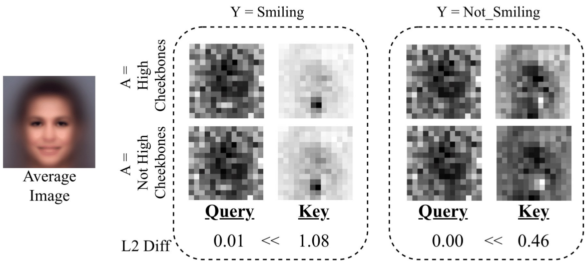

Figure 4: We show the average query and key activations for a fixed task label (Y=smiling or not smiling) and a fixed attribute label (A = High Cheekbones or not High Cheekbones) on the CelebA dataset after debiasing. The difference in query activations across attribute labels is reduced compared to the results observed before debiasing (Figure 1).

with TADeT (Table 1), validates our hypothesis that targeted debiasing focused on the query activations is an effective debiasing strategy for visual transformers.

Comparing CNN's and Transformers. We train a CNN on the same 3 settings described previously, and evaluate accuracy and fairness. As shown in Tab. 3, the baseline Transformer consistently performs significantly worse on both fairness metrics (Equalized Odds and Balanced Accuracy Difference). However, we refrain from concluding that Transformers are more biased than CNN's, because we see that Transformers also perform slightly worse in terms of accuracy measures (Balanced Accuracy and Standard Accuracy), which makes this an unfair comparison. As techniques develop to improve the accuracy of Vision Transformers to be commensurate with CNNs, we hope to address this question more conclusively.

	Y: Smiling **A:** High Cheekbones				**Y:** Wavy Hair **A:** Male				**Y:** Wavy Hair **A:** Wearing Lipstick			
Method	EO ↓	Δ BA (%) ↓	BA (%)↑	Acc (%)↑	EO ↓	Δ BA (%) ↓	BA (%)↑	Acc (%)↑	EO ↓	Δ BA (%) ↓	BA (%)↑	Acc (%)↑
CNN	14.66	2.69	88.15	93.06	16.71	8.08	77.99	82.20	15.64	6.41	78.75	82.20
Transformer	15.89	9.18	87.00	92.21	19.76	12.10	72.79	78.18	19.72	9.86	74.33	78.81

Table 3: Fairness and Accuracy metrics compared for a CNN and Transformer trained on the same task attribute-protected attribute combination. Y=Task. A=Protected Attribute. EO=Equalized Odds. ΔBA=Balanced Accuracy Difference. BA=Balanced Accuracy.

5 Conclusion

In this work, we perform the first benchmarking of several existing debiasing algorithms on visual transformers. We visualize the feature space learned by the transformer self-attention modules and find a significant portion of the bias is encoded in the query matrices. Using this information, we introduce TADeT, a targeted debiasing algorithm for transformers that aims to remove bias encoded within the query matrix.

Our method suffers from some limitations. First, we note that it is important for users to consider the application of their model before using our findings. In this work, we measure fairness using Equalized Odds and Balanced Accuracy Difference, but if one's intended use case does not align with these definitions of fairness, the results presented may not be applicable. Furthermore, all previous methods benchmarked in this paper, along with TADeT, require annotations for the protected attribute, which may not always be feasible.

Acknowledgments: This work was funded in part by Cisco Inc.

References

[1] Compas risk scales : Demonstrating accuracy equity and predictive parity performance of the compas risk scales in broward county. 2016.

[2] Shruti Bhargava and David Forsyth. Exposing and correcting the gender bias in image captioning datasets and models, 2019.

[3] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift., 2009.

[4] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[5] Emily Denton, Ben Hutchinson, Margaret Mitchell, Timnit Gebru, and Andrew Zaldivar. Image counterfactual sensitivity analysis for detecting unintended bias, 2020.

[6] 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. URL https://openreview.net/forum?id=YicbFdNTTy.

[7] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. Fairness in deep learning: A computational perspective, 2020.

[8] 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*, pages 214–226, 2012.

[9] Charles Elkan. The foundations of cost-sensitive learning, 2001.

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[11] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

[12] Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3819–3828, 2015.

[13] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey, 2021.

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[16] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[17] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

[18] Sungho Park, Dohyung Kim, Sunhee Hwang, and Hyeran Byun. Readme: Representation learning by fairness-aware disentangling method, 2020.

[19] Bernhard Schölkopf, John Platt, and Thomas Hofmann. *A Kernel Method for the Two-Sample-Problem*, pages 513–520. 2007.

[20] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11070–11078, 2020.

[21] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias. In *Domain adaptation in computer vision applications*, pages 37–55. Springer, 2017.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[23] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity, 2020.

[24] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5310–5319, 2019.

[25] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020.

[26] G. Weiss, Kate McCarthy, and Bibi Zabar. Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? In *DMIN*, 2007.

[27] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection, 2019. URL http://arxiv.org/abs/1902.11097.

[28] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[29] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1323. URL https://aclanthology.org/D17-1323.