Attention Aware Debiasing for Unbiased Model Prediction

Puspita Majumdar\textsuperscript{1,2}, Richa Singh\textsuperscript{2}, Mayank Vatsa\textsuperscript{2}
\textsuperscript{1}IIIT-Delhi, India \quad \textsuperscript{2} IIT Jodhpur, India
puspitam@iiitd.ac.in, \{richa, mvatsa\}@iitj.ac.in

Abstract

Due to the large applicability of AI systems in various applications, fairness in model predictions is extremely important to ensure that the systems work equally well for everyone. Biased feature representations might often lead to unfair model predictions. To address the concern, in this research, a novel method, termed as Attention Aware Debiasing (AAD) method, is proposed to learn unbiased feature representations. The proposed method uses an attention mechanism to focus on the features important for the main task while suppressing the features related to the sensitive attributes. This minimizes the model’s dependency on the sensitive attribute while performing the main task. Multiple experiments are performed on two publicly available datasets, MORPH and UTKFace, to showcase the effectiveness of the proposed AAD method for bias mitigation. The proposed AAD method enhances the overall model performance and reduces the disparity in model prediction across different subgroups.

1. Introduction

Artificial Intelligence (AI) systems trained using deep learning algorithms are increasingly used in real-world applications to support decisions in industry, government organizations, and law enforcement. Due to the high performance of AI systems, they are often deployed in high-stake applications for making predictions about individuals. However, AI systems have the risk of associating sensitive attributes (e.g., race, age) with the main task. Thus, a major concern of AI systems is their biased behavior against certain groups of individuals that are protected by law or ethics. Such biased behavior is often observed in facial analysis applications, including face recognition and attribute prediction \cite{1, 8, 19, 22}. For instance, commercial gender classifiers are shown to be biased towards lighter-skinned males in comparison to darker-skinned females \cite{7}. Apart from this, the biased behavior of facial processing technologies that mislabel Black faces as gorillas \cite{16} raises the concern towards fairness and trustability of AI systems. Due to the disparity in the performance of AI systems, some organizations have decided to minimize or ban the usage of facial analysis systems \cite{9}, while several others are still continuing the deployment and usage. Therefore, designing algorithms to mitigate bias and increase the fairness of deep models is of paramount importance.

Researchers have demonstrated that the distribution of training data plays a significant role in the performance of deep models \cite{6, 30}. An imbalance in training data distribution with respect to a particular subgroup (e.g., White and Asian are the subgroups of the sensitive attribute \textit{ethnicity}) leads to biased predictions \cite{4}. Therefore, several algorithms have been proposed to mitigate the effect of bias due to imbalanced data distribution either by over-sampling the under-represented subgroup or under-sampling the over-represented subgroup \cite{20, 21}. However, it is also highlighted that even models trained on balanced datasets amplify bias \cite{31}. Apart from this, researchers have shown that model training using conventional approaches has a direct correlation with bias in model predictions \cite{25}. For example, in conventional model training, the model automatically identifies and learns the features that maximize the overall model performance. However, in the learning process...
process, the model may learn the features related to the sensitive attributes that are not relevant to the main task, leading to biased predictions. Thus, it is important to incorporate some mechanism to ensure that the model learns the features related to the main task and ignores or suppresses the features related to the sensitive attributes during model training for mitigating bias in model prediction.

In this research, we have proposed a novel method, termed as Attention Aware Debiasing (AAD) method for bias mitigation. The proposed AAD method uses attention modules to enhance the model performance while suppressing the effect of sensitive attributes on model prediction for an unbiased outcome. In other words, the proposed method uses an attention mechanism to focus on important features and suppress the unnecessary ones to unlearn the model’s dependency on sensitive attributes. Figure 1 visually illustrates model training using conventional and the proposed AAD method. As shown in Figure 1(b), model training using the proposed AAD method utilizes the supervision of two attention modules to (i) focus on important features for the main task and (ii) suppress the features related to the sensitive attributes. For instance, consider face-based gender prediction as the main task with ethnicity as the sensitive attribute. During training, the model uses the Main Attention Module (MAM) for assigning higher weights to the features important for gender prediction and uses the Sensitive Attention Module (SAM) to assign lower weights to the features important for ethnicity prediction to learn unbiased features for the main task (gender prediction). The feature weights learned during training uses a multi-task network to perform the main task along with the sensitive attribute prediction, as shown in Figure 1(b).

The proposed AAD method improves the overall model performance by providing attention to the main task-specific features and decreases the bias in model prediction by unlearning the features related to the sensitive attribute. The attention modules used in the proposed AAD method can be added on top of a pre-trained model (feature extractor) to debias the feature representation for the main task. Therefore, only the attention modules and the sub-networks (for the main task and sensitive attribute prediction) of the multi-task network are updated during training, making the proposed method computationally efficient. The effectiveness of the proposed AAD method is demonstrated on two publicly available datasets, MORPH [27], and UTKFace [34] for bias mitigation in gender prediction.

2. Related Work

In the literature, researchers have studied the problem of bias and attempted to understand the presence of bias in model prediction followed by designing different algorithms for bias mitigation. An initial study is conducted in [7] to highlight the disparity in the performance of commercial gender classifiers when evaluated on the images of lighter and darker skin tones. By taking a step forward, Muthukumar et al. [22] analyzed the effect of varying skin tones on gender prediction and concluded that not only skin tone but the differences in ethnicity is a driving factor for the biased predictions. Denton et al. [11] examined the variations in classifier predictions due to small changes in facial characteristics using an image counterfactual sensitivity analysis framework. Towards understanding bias, Joo and Kärkkäinen [15] used an encoder-decoder network to generate face images with varying gender and ethnic groups. The generated images are used for measuring counter-factual fairness of commercial classifiers. Krishnan et al. [18] investigated the variations in the performance of gender classification algorithms across different gender-race groups. The authors analyzed the effect of different deep model architectures and imbalanced training sets on gender classification performance. The studies and analysis performed by the researchers helped to understand the sources of bias in model predictions and develop solutions for unbiased outcomes.

Several algorithms have been proposed to mitigate the effect of bias in model predictions. The majority of these algorithms learn unbiased feature representations either by using some pre-processing techniques on the input data or updating the feature representations of a model. These approaches require model training from scratch or re-training of a few convolutional layers. For instance, in 2018, Ryu et al. [29] proposed InclusiveFaceNet that used the transfer-learning approach for facial attribute detection. For mitigating soft biometrics-related bias, Das et al. [10] proposed a multi-task convolution neural network. The proposed multi-task network is used to perform a joint classification of race, gender, and age. Further, a joint learning and unlearning algorithm is proposed for removing bias in the feature representation of a network [2]. Kim et al. [17] proposed a regularization loss to unlearn the bias information by minimizing the mutual information between feature embedding and bias. A novel algorithm for bias mitigation in face detection using variational autoencoder is proposed by Amini et al. [3]. The proposed algorithm learned the latent structure within the dataset with respect to the ethnicity and gender of the subject to re-weight the samples during training. Recently, Nagpal et al. [23] proposed a novel filter-drop technique for efficient filter selection to mitigate the effect of bias in model prediction. On the other hand, Roh et al. [28] proposed a technique to adaptively select mini-batches during training for improving model fairness. A technique, termed as diversity block, is proposed in [24] to de-bias pre-trained models. Here, the authors attached the diversity block to an existing pre-trained model and trained it separately on a small training data against which the pre-trained model is biased. Another work on mitigating bias...
3. Proposed Method

In this research, we have proposed an approach that uses network attention to perform bias-invariant and efficient prediction by suppressing the effect of the sensitive attribute on model prediction. Figure 2 shows the block diagram of the model training using the proposed AAD method.

3.1. Attention Aware Debiasing (AAD) Method

Our AAD method is inspired by the work [32]. Attention networks are used in the literature to enhance the overall model performance [5, 14]. However, unlike the previous works on attention networks, the proposed AAD method uses an attention mechanism for bias mitigation by unlearning the features related to the sensitive attribute. The attention modules used in our proposed AAD method are added on top of a pre-trained model (feature extractor). Figure 3 shows the framework of the attention modules.

Let $\phi$ be a multi-task convolutional neural network with two tasks: main task and sensitive attribute prediction. Consider $x_i$ as an input image which is given as input to the multi-task network $\phi$. Let $v_i$ be a $d \times 1$ dimensional feature vector (corresponding to image $x_i$) obtained after flattening the output of the last convolutional layer of the multi-task network $\phi$. The feature vector $v_i$ is given as input to the main and sensitive attention modules to learn the features important for the main task and suppress the features important for sensitive attribute prediction. The attention module is a multi-layer perceptron (MLP) with a single hidden layer followed by the sigmoid function (Figure 3). Let $A_m$ and $A_s$ represent the main and sensitive attention modules, respectively.

$$o_{m_i} = A_m(v_i)$$  \hspace{1cm} (1) \\
$$o_{s_i} = A_s(v_i)$$  \hspace{1cm} (2) \\

where, $o_{m_i}$ and $o_{s_i}$ represent the outputs of the main and sensitive attention modules, respectively. The feature vector $v_i$ is combined with the outputs of the attention modules to obtain the updated feature vectors for each task. The updated feature vectors are then used for the main task and sensitive attribute prediction. The updated feature vectors $v_{m_i}$ for the main task and $v_{s_i}$ for the sensitive attribute prediction are obtained using the following equations:

$$v_{m_i} = o_{m_i} \otimes (1 - o_{s_i}) \otimes v_i$$  \hspace{1cm} (3) \\
$$v_{s_i} = o_{s_i} \otimes v_i$$  \hspace{1cm} (4)
where, $\otimes$ denotes element-wise multiplication. In Equation 3, $o_m$ weigh the features important for the main task and $(1 - o_s)$ suppress the sensitive features that are not relevant to the main task. In other words, during training, the network tries to learn important features for the main task independent of the sensitive attribute. It is important to note that a high-performing sensitive attribute predictor is required to provide better supervision to suppress the features related to the sensitive attribute. Thus, $o_s$ weigh the features important for sensitive attribute prediction to ensure enhanced performance of sensitive attribute predictor (Equation 4).

As shown in Figure 2, a multi-task network is learned for the main task along with the sensitive attribute prediction. Let $\phi_m$ represent the sub-network for the main task, which takes feature vector $v_m$, as input and outputs the probability vector for a class $m_k$. The output of $\phi_m$ for image $x_i$ is represented as:

$$P(m_k|x_i) = \phi_m(v_m)$$  \hspace{1cm} (5)

where, $P(m_k|x_i)$ is the probability of predicting image $x_i$ to $m_k$. The loss function for the main task is represented as:

$$L_{m_i} = -\sum_{\forall k} m_k \log P(m_k|x_i)$$  \hspace{1cm} (6)

Let $\phi_s$ represent the sub-network for sensitive attribute prediction, which takes feature vector $v_s$, as input and outputs the probability vector for sensitive attribute class $s_n$. The output of $\phi_s$ for image $x_i$ is represented as:

$$P(s_n|x_i) = \phi_s(v_s)$$  \hspace{1cm} (7)

where, $P(s_n|x_i)$ is the probability of predicting image $x_i$ to $s_n$. The loss function for sensitive attribute prediction is represented as:

$$L_{s_i} = -\sum_{\forall n} s_n \log P(s_n|x_i)$$  \hspace{1cm} (8)

The final loss function used to train the multi-task network via the attention modules is written as follows:

$$L = \sum_i (\lambda L_{m_i} + L_{s_i})$$  \hspace{1cm} (9)

where, $\lambda$ is a hyper-parameter. The loss $L$ simultaneously improves the overall model performance and reduces bias in model prediction.

### 3.2. Bias-Invariant Prediction

The proposed AAD method helps to learn unbiased representations for bias-invariant prediction. AAD method ensures that the features encoding the sensitive attribute are not used for the main task. Figure 4 shows the block diagram of the network during testing. It is important to note that the network is a uni-task network during testing. As shown in Figure 4, the sub-network for the main task is used for the final prediction, which uses both the attention modules to obtain unbiased feature representations for prediction.

### 4. Experimental Setup

The performance of the proposed AAD method is evaluated for the task of gender prediction. A gender prediction model classifies an input image into male or female. Two publicly available datasets are used to perform the experiments. The following discusses the details of the datasets with the corresponding protocols, implementation details, and evaluation metrics.

#### 4.1. Datasets and Protocols

Experiments are performed on the following datasets to evaluate the performance of the gender prediction model across two sensitive attributes: ethnicity and age.

**MORPH dataset (Album-2)** [27] contains more than 54K images of 13K subjects. The dataset is pre-labeled with two genders (male and female), six ethnicities (White, Black, Hispanic, Indian, Asian, and Other), and age (ranging from 16 to 77 years). The dataset is partitioned with non-overlapping subjects in the training and testing sets. The training set contains 70% subjects while the testing set contains 30% subjects. For the experiments, images belonging to the White and Black ethnicity are used. Also, we considered images below 36 years as Young and the rest as Old.

**UTKFace dataset** [34] consists of more than 20,000 face images. The dataset is pre-labeled with two genders.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Main Task</th>
<th>Bias Mitigation Across</th>
</tr>
</thead>
<tbody>
<tr>
<td>MORPH</td>
<td>Gender Prediction</td>
<td>Ethnicity ($E_W$, $E_B$), Age ($A_Y$, $A_O$)</td>
</tr>
<tr>
<td>UTKFace</td>
<td>Gender Prediction</td>
<td>Ethnicity ($E_W$, $E_A$), Age ($A_Y$, $A_O$)</td>
</tr>
</tbody>
</table>

### Table 1. Details of the experiments for bias mitigation in gender prediction across different sensitive attributes.

Figure 4. Illustrating the utilization of the proposed Attention Aware Debiasing Method (AAD) for the main task during testing. The sub-network for the sensitive attribute prediction is not used for the final prediction.
(male and female), five ethnicities (White, Black, Asian, Indian, and Others), and age (ranging from 0 to 116 years). We partitioned the dataset into disjoint training and testing sets with 70% images in the training set and 30% in the testing set. For the experiments, images belonging to the White and Asian ethnicity are used. Also, we considered images below 26 years as Young and the rest as Old.

For both the datasets, the training and testing partitions are balanced with respect to gender, ethnicity, and age subgroups. The ethnicities are denoted by \(E_W\), \(E_B\), and \(E_A\) for White, Black, and Asian, respectively. Similarly, the age groups are denoted by \(A_Y\) and \(A_O\) for Young and Old, respectively. The details of the experiments are summarized in Table 1. Figure 5 shows some sample images of the datasets.

**4.2. Implementation Details**

Experiments are performed using LightCNN-29 [33] architecture. The model weights are initialized with those learned on the MS-Celeb-1M dataset [13]. As shown in Figure 2, the attention modules are added after the final convolutional layer. The attention modules consist of two separate multi-layer perceptrons (Figure 3). The multi-layer perceptron of both the attention modules consists of three layers of dimensions 256, 128, and 256, respectively. The attention modules are followed by the sub-networks for the main task and sensitive attribute prediction. Both the sub-networks consist of two dense layers of dimensions 128 and 64, respectively. Each layer is followed by ReLU activation.

The weights of the convolutional layers are kept frozen (treated as a pre-trained model for feature extraction), and only the attention modules and the sub-networks for the main task and sensitive attribute prediction are trained. The multi-task network is trained for 10 epochs, using the Stochastic Gradient Descent optimizer with 0.001 learning rate. Momentum is set to 0.9 and batch size to 50. During the experiment, the \(\lambda\) parameter is set to 6 and 2 for the MORPH and UTKFace datasets, respectively. Code is im-

**Table 2. Performance of the proposed and existing methods (%) for gender prediction on the MORPH dataset across different ethnicity and age groups.**

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy ↑</th>
<th>DoB ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(E_W)</td>
<td>(E_B)</td>
</tr>
<tr>
<td>Traditional</td>
<td>92.51</td>
<td>94.20</td>
</tr>
<tr>
<td>MTL [10]</td>
<td>92.37</td>
<td>94.63</td>
</tr>
<tr>
<td>DB [24]</td>
<td>93.50</td>
<td>92.09</td>
</tr>
<tr>
<td>Proposed</td>
<td>96.04</td>
<td>94.77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy ↑</th>
<th>DoB ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(A_Y)</td>
<td>(A_O)</td>
</tr>
<tr>
<td>Traditional</td>
<td>91.66</td>
<td>95.05</td>
</tr>
<tr>
<td>MTL [10]</td>
<td>91.38</td>
<td>95.62</td>
</tr>
<tr>
<td>DB [24]</td>
<td>91.81</td>
<td>94.49</td>
</tr>
<tr>
<td>Proposed</td>
<td>93.22</td>
<td>95.19</td>
</tr>
</tbody>
</table>
implemented in PyTorch. All the experiments are performed on Nvidia GeForce GTX 1080 Ti.

4.3. Evaluation Metrics

Experimental results are reported using performance and bias evaluation metrics. We have used overall and class-wise classification accuracy for performance evaluation. For measuring bias in model prediction, we have used Degree of Bias (DoB) [12], which measures the standard deviation of classification accuracy across different subgroups of a sensitive attribute. A lower value of DoB indicates low bias in model prediction.

5. Results and Analysis

The performance of the proposed AAD method is evaluated for bias mitigation in gender prediction across different ethnicity and age groups. The proposed AAD method is compared with traditional and multi-task model training (MTL) methods [10]. In traditional model training, the model is trained only for the task of gender prediction (without sensitive attribute predictor). On the other hand, in multi-task model training, the model is trained for both gender prediction and sensitive attribute prediction (similar to [10]). The traditional model training method is used for comparison to highlight the drawbacks of conventional model training approaches that lead to biased predictions. Further, the comparison with the multi-task model training method is done with the aim of analyzing the effectiveness of the attention modules for unbiased model predictions. To compare the proposed AAD method with existing bias mitigation algorithms to de-bias pre-trained models, we have compared our method with Diversity Block (DB) technique [24]. As mentioned in the related work section, the diversity block technique is a recently proposed technique used to mitigate bias in pre-trained models.

Table 2 shows the results of gender prediction across different sensitive attributes on the MORPH dataset. It is observed that the proposed AAD method improves the overall model performance and reduces bias in model prediction compared to the existing methods. For instance, the proposed AAD method increases the overall classification accuracy from 93.36% to 95.40% and reduces the DoB from 0.84% to 0.63% compared to the traditional method across different ethnicity. It is also observed that the multi-task model training (MTL) method increases the overall model performance compared to the traditional model training method. However, the disparity in model performance also increases across different subgroups. On the other hand, the proposed AAD method that uses the attention modules in a multi-task learning setup reduces the disparity in model performance across different subgroups. This highlights the efficacy of the attention modules to unlearn the features related to the sensitive attributes for gender prediction. For further analysis, we have compared the features extracted from the last convolutional of the model with the updated features obtained after the attention modules using t-Distributed Stochastic Neighbor Embedding (t-SNE) visualizations. Figure 6 shows the t-SNE visualizations of both features extracted after the last convolutional layer of the model. Figure 6 also shows the visualization of the updated feature representation obtained after the attention modules corresponding to the model trained using the proposed AAD method to unlearn the ethnicity-related features during gender prediction.

Table 3. Performance of the proposed and existing methods (%) for gender prediction on the UTKFace dataset across different ethnicity and age groups.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy ↑</th>
<th>DoB ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EW</td>
<td>EA</td>
</tr>
<tr>
<td>Traditional</td>
<td>81.50</td>
<td>87.00</td>
</tr>
<tr>
<td>MTL [10]</td>
<td>81.37</td>
<td>87.50</td>
</tr>
<tr>
<td>DB [24]</td>
<td>81.88</td>
<td>83.88</td>
</tr>
<tr>
<td>Proposed</td>
<td>85.37</td>
<td>88.12</td>
</tr>
</tbody>
</table>

Figure 6. t-SNE visualizations of the 256-dimensional feature representation of the testing set corresponding to the MORPH dataset for gender prediction. (a) Shows the visualization of the feature representation obtained after the last convolutional layer of the model. (b) Shows the visualization of the updated feature representation obtained after the attention modules corresponding to the model trained using the proposed AAD method to unlearn the ethnicity-related features during gender prediction.

4138
Table 4. Confusion matrix (%) for gender prediction (Male as 'M' and Female as 'F') across different ethnicity and age groups on the UTKFace dataset.

<table>
<thead>
<tr>
<th>Across Ethnicity</th>
<th>Across Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>E_W</td>
</tr>
<tr>
<td>Ground Truth</td>
<td>Predicted</td>
</tr>
<tr>
<td>Traditional</td>
<td>M 77.75</td>
</tr>
<tr>
<td></td>
<td>F 14.75</td>
</tr>
<tr>
<td>Proposed</td>
<td>M 83.50</td>
</tr>
<tr>
<td></td>
<td>F 12.75</td>
</tr>
</tbody>
</table>

Figure 7. Sample images of the UTKFace dataset, misclassified by the proposed AAD method. Large variations in pose, illumination, resolution, and occlusion make the problem more challenging.

shows that the Diversity Block (DB) technique reduces bias in model prediction while it compromises the overall model performance. On the other hand, the proposed AAD method demonstrates enhanced overall performance with reduced effect of bias in model prediction. This showcases the efficacy of the proposed method.

The results on the UTKFace dataset are summarized in Table 3. The proposed AAD method outperforms existing model training methods. For instance, the proposed AAD method achieves 85.81% accuracy and 1.69% DoB for gender prediction across different age groups. This shows that the proposed AAD method that uses an attention mechanism for learning unbiased feature representation is effective for bias mitigation. The confusion matrix for gender prediction across different sensitive attributes on the UTKFace dataset is shown in Table 4. The proposed AAD method shows high performance for both the classes (male and female). Figure 7 shows some sample images misclassified by the proposed AAD method. From Figure 7, it is observed that most of the images have large variations in pose, resolution, and illumination. Some images have partial occlusion due to eyeglasses and caps. Gender prediction becomes difficult in the presence of these covariates, leading to misclassification of these images.

On comparing the results of Tables 2 and 3, it is found that the overall model performance is higher on the MORPH dataset compared to the UTKFace dataset using existing and the proposed methods. This is due to the fact that the images in the MORPH dataset are captured in constrained environmental settings, while the images in the UTKFace dataset are captured in unconstrained environmental settings with large variations in pose, resolution, illumination, and degree of occlusion. This showcases the challenges of unconstrained gender prediction.

6. Ablation Study

Experiments are performed by ablating the Sensitive Attention Module (SAM) along with the sub-network for sensitive attribute prediction to analyze their role towards mitigating the effect of bias in gender prediction. Therefore, in this experiment, the model with the Main Attention Module (MAM) and the sub-network for the main task are trained only for gender prediction. It is important to note that the model does not get the supervision of the SAM to suppress the feature related to the sensitive attribute. Thus, the output of the MAM is combined only with the features of the last convolutional layers to weigh the features important for gender prediction.

Figure 8 shows the overall classification accuracy and DoB of the ablated model for gender prediction on both datasets. Comparison is performed with the proposed AAD method. It is observed that the disparity in model performance across different subgroups of a sensitive attribute is higher for the ablated model. For instance, the DoB of the ablated model is 3.81% across different age groups on the UTKFace dataset, which is 2.12% higher than the proposed AAD method. This showcases the importance of
SAM and the sub-network for sensitive attribute prediction towards unlearning the model’s dependency on the sensitive attribute. The supervision of the SAM is important to learn unbiased feature representation for gender prediction. Further, on comparing the classification performance, it is found that the ablated model achieves almost equal accuracy compared to the proposed AAD method. Here, the MAM focuses on the features important for gender prediction, thereby enhancing the overall classification accuracy. This shows the importance of attention networks towards improving the overall model performance.

7. Conclusion and Discussion

The advancements in deep learning techniques and the availability of large-scale datasets have led to the development of sophisticated AI systems that achieve high accuracy for various classification/prediction tasks. Thus, AI systems are widely used and deployed in various real-world applications that affect every aspect of our lives. However, several incidents have highlighted the biased behavior of AI systems with respect to protected groups, raising concern towards the trustability and dependability of these systems. Therefore, fairness in AI systems is of paramount importance for unbiased model predictions.

Deep models automatically learn the features from the input data that maximize the model performance. However, in the learning process, the model may learn biased features that favor or disfavor a particular subgroup, leading to unfair model predictions. Therefore, it is crucial to design a mechanism for learning unbiased feature representations for fair outcomes. This research presents a solution to the problem using the proposed Attention Aware Debiasing (AAD) method. The proposed AAD method uses an attention mechanism to learn unbiased feature representations by unlearning the model’s dependency on the sensitive attribute. The supervision provided by the attention modules is utilized to focus on the features relevant for the main task and suppress the features related to the sensitive attribute.

The efficacy of the proposed AAD method is shown for the task of gender prediction. Experimental results highlight that the proposed AAD method is able to mitigate bias in model prediction and enhance the overall model performance. Further, the attention modules used in the proposed AAD method can be added on top of a wide variety of pre-trained models to perform various tasks in different domains. In the current experimental setup, the proposed AAD method is used for bias mitigation in a single task (gender prediction). As a part of future work, the proposed AAD method can be extended to learn unbiased feature representations for multiple tasks to mitigate bias due to various sensitive attributes.

Acknowledgements

P. Majumdar is supported by DST Inspire Ph.D. Fellowship. M. Vatsa is partially supported through Swarnajayanti Fellowship. This research is also partially supported by Facebook Ethics in AI award.

References


[6] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embed-


[34] Ziheng Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In IEEE Conference on Computer Vision and Pattern Recognition, pages 5810–5818, 2017. 2, 4, 5