2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

InsideBias: Measuring Bias in Deep Networks and Application to Face Gender Biometrics

Ignacio Serna, Alejandro Peña, Aythami Morales, Julian Fierrez
School of Engineering, Universidad Autonoma de Madrid, Spain
{ignacio.serna, alejandro.penna, aythami.morales, julian.fierrez}@uam.es

***Abstract*—This work explores the biases in learning processes based on deep neural network architectures. We analyze how bias affects deep learning processes through a toy example using the MNIST database and a case study in gender detection from face images. We employ two gender detection models based on popular deep neural networks. We present a comprehensive analysis of bias effects when using an unbalanced training dataset on the features learned by the models. We show how bias impacts in the activations of gender detection models based on face images. We finally propose InsideBias, a novel method to detect biased models. InsideBias is based on how the models represent the information instead of how they perform, which is the normal practice in other existing methods for bias detection. Our strategy with InsideBias allows to detect biased models with very few samples (only 15 images in our case study). Our experiments include 72K face images from 24K identities and 3 ethnic groups.**

I. Introduction

Artificial Intelligence (AI) algorithms have an increasingly growing role in our daily lives. These algorithms influence now many decision-making processes affecting people's lives in many important fields, e.g. social networks, forensics, health, and banking. For example, some companies already use AI to predict credit risk, and some US states run prisoner details through AI systems to predict the likelihood of recidivism when considering parole [1].

Face recognition algorithms are good examples of recent advances in AI. During the last ten years, the accuracy of face recognition systems has increased up to 1000x (it is probably the biometric technology with the greatest investment nowadays). These face recognition algorithms are dominated by Deep Neural Network architectures, which are trained with huge amounts of data with little control over what is happening during training (focused on performance maximization). As a result, we have algorithms with excellent performance but quite opaque.

This trend in AI (excellent performance + low transparency) can be observed not only in face biometrics, but also in many other AI applications as well [2]. At this point, and despite the extraordinary advances in recognition performance, factors such as the lack of transparency, discrimination, and privacy issues are limiting many AI practical applications. As an example of these increasing concerns, in May 2019, the Board of Supervisors of San Francisco banned the use of facial recognition software by the police and other agencies, and many others are considering or already have enacted legislation [3].

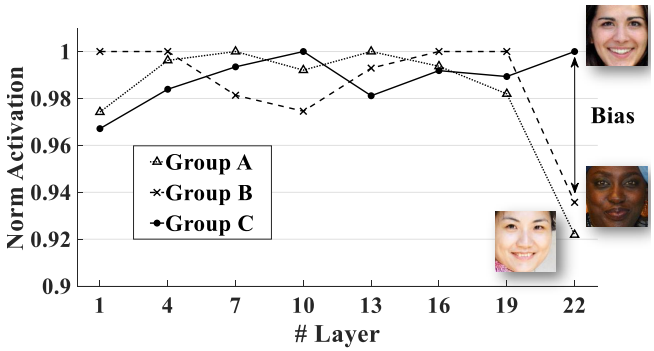


Fig. 1. Normalized overall activation observed through layers of a biased ResNet model trained for facial gender recognition. Overall activation for the images of Group A, Group B and Group C (See Section IV). The arrow on the right highlights the difference in activations obtained between well-represented and poorly-represented groups.

The number of published works pointing out the biases in the results of face recognition algorithms is large [4–7]. Among these works, a vast majority analyzes how biases affect the performances obtained for different demographic groups [3, 4, 6, 8–17]. However, only a limited number of works analyze how biases affect the learning process of these algorithms. In this work we use feature visualization techniques of deep models to generate a better understanding of biased learning. We will analyze two gender recognition models based on face images and how the ethnicity bias affects the learning process.

Given an input, the activations of the different neurons inside a neural network define the output of the network. Fig. 1 shows the difference in the overall activation of a gender recognition network for input face images from three different demographic groups. The figure shows the overall activations at different layers for face images of a well-represented demographic group (caucasian female in this example), and poorly-represented demographic groups (black and asian females in this example). Well-represented group refers to a demographic group prioritized during the training phase of the network with many more samples compared to poorly-represented groups. The overall activation is measured for each image in each layer and then averaged across images of the same demographic group. Strong activations are usually related with highly discriminative features [18]. In the figure, for the well-represented group we can observe how

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9412443

the activations of the gender classification network are high for the first layers (focused on characterizing textures and colors) increasing slightly for the last layers (focused on the high level features related to the gender classification task). In comparison, for the poorly-represented group the activations decrease significantly towards the final layers. These low activations may be the cause of biased performances that lead to unfair discrimination.

The contributions of this work are twofold:

- Comprehensive analysis of biased learning focused on the effects over the activation level of learned features in two public databases; demonstrating the latent correlations between bias, activation, and recognition performance.
- We propose InsideBias, a novel bias detection method based on the analysis of the filters activation of deep networks. InsideBias allows to detect biased models with very few samples (only 15 in our experiments), which makes it very useful for algorithm auditing and discrimination-aware machine learning.

The rest of the paper is structured as follows: Section II summarizes related works in this area. Section III introduces the formulation of the problem and the method proposed in this work. Section IV presents the experimental framework including databases, models, and experimental protocols. Section V analyzes experimentally how bias affects different models, and presents the results obtained by the proposed bias detection method InsideBias. Finally, Section VI summarizes the main conclusions.

II. Related Work

A. Demographic biases in biometrics and AI at large

The study carried out in [19] examined three commercial face detection algorithms. It revealed the discriminatory potential of these technologies and attracted the interest of the academia and industry, reaching widespread impact even in mass media. This presence in the general press comes with a risk: possible misinterpretation due to excessive generalization from the particular experiments reported in that paper (based on three particular face detection systems) to face biometrics in general, and AI at large.

The existence of undesired demographic bias and algorithmic discrimination in AI depend heavily on many factors such as: learning architecture, training strategy, target problem, and performance criteria [20]. The present work, as others recently [17], generates a better understanding of such factors: 1) as input knowledge for more informed bias analyses, and 2) to generate tools to deal with such undesired biases in practical problems.

As a representative example in this line of work examining demographic biases in AI, Nagpal et al. conducted a series of experiments to verify if deep neural networks in facial recognition systems encode some type of race-specific information [7], and found that in models trained with different races, different discriminative regions contribute to the decision.

The same algorithm can behave very differently when tested on different groups of samples. For example, [3] demonstrated that same face recognition algorithm performed much worse for dark-skinned people than for light-skinned people, given images from some cameras, but performed better for images from other cameras.

Turning to the specific case of face recognition, the work [14] proves that bias is not linked exclusively to demographic factors. Moreover, it discusses how sensational headlines written by non-expert people skew the information around biases in AI.

B. Looking inside Neural Networks

As soon as the rebirth of Neural Networks happened in the past decade, researchers have tried to generate a better understanding of the representations learned by neural models.

Erhan et al. proposed an approach to visualize the hidden layers [21]. Zeiler and Fergus applied a deconvolution algorithm to see the activity within the model [18], and Simonyan et al. generated the representative image of a class by maximizing the class scores [22]. Yosinski et al. visualized the activations of each neuron when processing an image [23]. Selvaraju et al. introduced an algorithm that visually highlights a network's decision by computing the gradient of the class score with respect to the input image [24]. In a similar line of work, Nguyen et al. created synthetic images that maximally activated each neuron [25], and Olah et al. explored which neurons are activated in different regions of the image [26, 27].

In addition, there are other variants and improvements of the methods indicated above, like the ones that identify characteristics encoded by the network relevant to the task [28], or other ways to intervene certain neurons in order to see the effect they have [29].

Inspired by the literature, in the present work we look at the raw activations of the neurons in presence of demographic biases in gender classification algorithms.

III. Measuring Bias in Deep Networks: InsideBias

A. Formulation of the problem

Let's begin with notation and preliminary definitions. Assume $\mathbf{I}$ is an input sample (e.g. face image) of an individual. That sample $\mathbf{I}$ is assumed to be useful for task T, e.g., face authentication or gender recognition. That sample is part of a given dataset $\mathcal{D}$ (collection of multiple samples from multiple subjects) used to train a model defined by its parameters $\mathbf{w}$. We also assume that there is a goodness criterion G_T on that task T maximizing some performance function f_T in the given dataset $\mathcal{D}$ in the form:

$$G_T(\mathcal{D}) = \max_{\mathbf{w}} f_T(\mathcal{D}, \mathbf{w}) \tag{1}$$

On the other hand, the individuals in $\mathcal{D}$ can be classified according to two demographic criteria (without loss of generality, we can have more criteria): $d = 1 \equiv \textit{Gender} \in \{\textit{Male}, \textit{Female}\}$ and $d = 2 \equiv \textit{Ethnicity} \in \{A, B, C\}$. We assume that all classes are well represented in dataset $\mathcal{D}$, i.e., the number of samples of each class for all criteria in $\mathcal{D}$ is

significant. $\mathcal{D}_d^k \subset \mathcal{D}$ represents all the samples corresponding to class k of demographic criterion d.

In our experiments, the goodness criterion $G_T(\mathcal{D})$ is defined as the performance of a gender recognition algorithm (T = *Gender Recognition*) on the dataset $\mathcal{D}$. During the experiments, we study how the criterion $d = 1 \equiv$ *Ethnicity* affects the internals of an algorithm focused on recognition of a different criterion $d = 2 \equiv$ *Gender*.

B. Bias estimation with InsideBias

While most of the literature is focused on estimating bias through performance between different datasets [17], with InsideBias we propose a novel approach based on the activation levels[1] within the network for different datasets $\mathcal{D}_d^k$ from different demographic groups.

Convolutional Neural Networks are composed by a large number of stacked filters. These filters are trained to extract the richest information for a predefined task (e.g. digit classification or gender recognition). These filters are activated as an input (e.g. an image) goes through the network. Stronger activations are usually related to the detection of highly discriminative features [18].

Without loss of generality, we present InsideBias for Convolutional Neural Networks (CNNs). Similar ideas are extensible to other neural learning architectures. In a convolutional layer of a CNN, the previous layer's feature maps are convolved with the filters (also known as kernels) and put through the activation function to form the output feature map. The output $\mathbf{A}^{[l]}$ of layer l consists of $m^{[l]}$ feature maps of size $n_1^{[l]} \times n_2^{[l]}$, where $m^{[l]}$ is the number of filters at layer l. The i^{th} feature map in layer l denoted as $\mathbf{A}_i^{[l]}$ is computed as:

$$\mathbf{A}_i^{[l]} = g^{[l]} \left(\sum_{j=1}^{m^{[l-1]}} \mathbf{f}_{ij}^{[l]} * \mathbf{A}_j^{[l-1]} + \mathbf{b}_i^{[l]} \right) \tag{2}$$

where $g^{[l]}$ denotes the activation function of the l^{th} layer, $*$ is the convolutional operator, $\mathbf{b}_i^{[l]}$ is a bias vector at layer l for the i^{th} feature map, and $\mathbf{f}_{ij}^{[l]}$ is the filter connecting the j^{th} feature map in layer $(l-1)$ with i^{th} feature map in layer l. The average activation of the i^{th} feature map at layer l is calculated as:

$$\overline{A_i^{[l]}} = \frac{1}{n_1^{[l]} \cdot n_2^{[l]}} \sum_{x=1}^{n_1^{[l]}} \sum_{y=1}^{n_2^{[l]}} A_i^{[l]}(x, y) \tag{3}$$

where (x, y) are the spatial coordinates of the output $A_i^{[l]}$. The activation, $\lambda^{[l]}$, is calculated as the maximum of $\overline{A_i^{[l]}}$ for all feature maps in the layer l:

$$\lambda^{[l]} = \max_i \left(\overline{A_i^{[l]}} \right) \tag{4}$$

We have evaluated both the average and the maximum, but the maximum resulted in a better estimator. Our intuition is that the maximum is related to highly discriminant patterns (high activations) and this is highly correlated to bias effects.

Filters tend to be different between networks trained differently, even if the networks have the same architecture, and even if they have been trained with the same data. The reason for this is that if the initialization to solve Eq. 1 (which is done iteratively) is different, since the solution space is very large [30], the solution of that equation will typically be a local minimum and will also depend on the particular training configuration. So to be able to compare the activations of different models we propose to normalize them:

$$\lambda'^{[l]} = \frac{\lambda^{[l]}}{\max_l \lambda^{[l]}} \tag{5}$$

The *Activation Ratio* $\Lambda_d^{[l]}$ for demographic criterion d (e.g., *Ethnicity* in our experiments) is then calculated as the ratio between the activation obtained for the group with the lowest $\lambda^{[l]}$ and the group with the highest $\lambda^{[l]}$:

$$\Lambda_d^{[l]} = \frac{\min_k \lambda^{[l]}(\mathcal{D}_d^k)}{\max_k \lambda^{[l]}(\mathcal{D}_d^k)} \tag{6}$$

InsideBias uses this *Activation Ratio* to detect biased models. A model will be considered biased if the *Activation Ratio* is smaller than a threshold τ. When analyzing the bias in this way we recommend looking at the final layers, similar to the initial example in Fig. 1.

IV. Experimental Framework

We start our bias experiments by studying how bias influence the data-driven learning process. For that we trained different architectures for two tasks: digit recognition and gender classification. To better understand its effects and the relationship with the activations, we will analyze the results of training with and without biased training data.

A. Databases

We have evaluated our approach on two different datasets:

1) Colored MNIST: Inspired in the experiment proposed in [31], we introduced bias in the form of colors into the MNIST dataset [32]. We used the 3 RGB colors (Red, Green, and Blue). Each digit in the training set (60K samples) was colored according to a highly biased distribution (i.e. 90% of the samples colored with a primary color and 10% with the remaining two colors). The digits in the test set (10K samples) were colored with an uniform distribution (i.e. 33-33-33). The goal is to analyze the impact of the color information in the learning process of the digit recognition model.

2) DiveFace: The second database used is the DiveFace dataset [33]. DiveFace contains annotations equally distributed among six classes related to gender and ethnicity. There are 24K identities (4K per class) and 3 images per identity for a total number of images equal to 72K. Users are grouped according to their gender (male or female) and three categories related with ethnic physical characteristics:

- **Group A**: people with ancestral origin in Japan, China, Korea, and other countries in that region.

[1]We refer to activation level as the output of the activation function of each neuron.

Fig. 2. Example images for the demographic **Groups** (**A, B** and **C**) used in our experiments.

- **Group B**: people with ancestral origins in Sub-Saharan Africa, India, Bangladesh, Bhutan, among others.
- **Group C**: people with ancestral origins in Europe, North-America, and Latin-America (with European origin).

Fig. 2 shows 15 face images examples from the three ethnic groups of DiveFace. Note that all images show similar pose, illumination, and quality. These images obtain very high confidence values in the gender recognition algorithms of this paper (i.e. confidence scores greater than 99%). The confidence score is the output of the network, it indicates the probability of belonging to one class (in our case the class of being a man or a woman).

Note that these are heterogeneous groups that include people of different ethnicities. We are aware of the limitations of grouping all human ethnic origins into only three categories. According to studies, there are more than 5K ethnic groups in the world. We have classified them into only three groups in order to maximize differences between classes. Automatic classification algorithms based on these three categories show performances of up to 98% accuracy [6].

B. Learning architectures

We employed two popular state-of-the-art image recognition architectures based on Convolutional and Residual layers. These architectures have been chosen as examples of standard models employed in face attribute detection algorithms [34, 35]:

Network 1 (VGG architecture [36]): The network is composed of eight convolutional layers followed by two fully connected layers with dropout. We use the ReLU (Rectified Linear Unit) activation function in all hidden layers, and a softmax activation function for the output layer (with two output units). This network comprises more than 660K parameters and its input is 120×120 for the gender recognition model and 28×28 for the digit recognition model.

Network 2 (ResNet architecture [37]): The network consists of three building blocks and a fully connected layer with softmax activation for the output layer (with two output units). Each building block is composed of convolutional layers. The big difference with the VGG architecture is the shortcut connections: within each block there is a shortcut connection that performs a convolution and bypasses a certain number of convolutional layers. This network comprises more than 370K parameters and its input is 120×120.

C. Experimental protocol

1) Colored MNIST Protocol: We have trained 31 digit recognition models using the VGG architecture:

- *Biased Models:* to analyze the impact of biased training data on the learning process, we decided to apply a highly biased color distribution in each of the ten digits of MNIST dataset. We defined three possible colors for a digit (Red, Green, and Blue). The highly biased distribution means that 90% of the training samples of a digit are colored with a primary color (e.g. Red), and the other 10% with the secondary colors (e.g. Green and Blue). Of the remaining 9 digits, all their training samples are colored with one of the secondary colors. This process is repeated for the ten digits and the three colors, resulting in 30 different models with 30 different biases.
- *Unbiased Model:* one model is trained with uniform color distribution (33%-33%-33%).

The color in the test set is assigned uniformly 33%-33%-33%. The goal of this experimental protocol is to analyze the relationship between bias, performance, and activations.

2) Gender Recognition Protocol: We have trained four models of each of the two chosen learning architectures, according to three different experimental protocols:

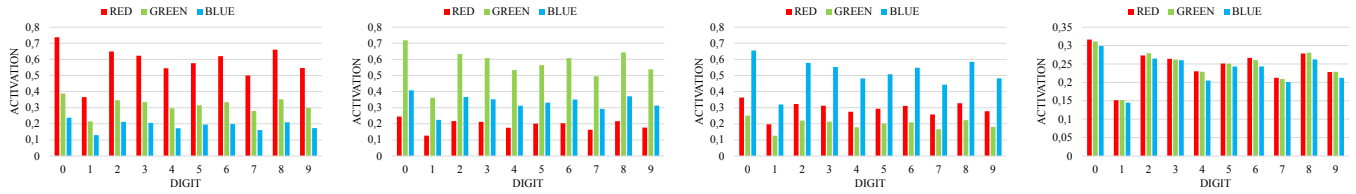
- *Biased Models*: the models in this experiment are trained with 18K images giving priority to one ethnic group with 90% of the images as opposed to the other two ethnic groups which are 5% and 5% respectively (all divided equally between men and women). This experiment is repeated 3 times giving preference to each ethnic group. Therefore, 3 independent models per network architecture are trained.
- *Unbiased Model* (trained with limited data): the models in this experiment are trained with 18K images (same number of images than biased models), 6K from each ethnic group, divided in half between men and women.

All models are evaluated with 18K images distributed equally among all three ethnic groups. None of the validation users have been used for training; i.e., it is an independent set.

V. Experimental Results

A. Role of the biased data

1) Colored MNIST: Fig. 3 shows the average activation $\lambda^{[l]}$ for the different models trained using the colored MNIST dataset and the protocol described in Section IV-C. The results demonstrate the correlation between bias, performance and activations. The poor performance obtained by highly biased models is not surprising. The biased distribution of color introduced in the training set decreases the performance of the network when testing in a set with uniform distribution of color. The experiment with the colored MNIST suggests that



(a) Red *Biased* (36.86%) (b) Green *Biased* (36.98%) (c) Blue *Biased* (36.35%) (d) *Unbiased* (97.87%)

Fig. 3. Average $\lambda^{[l]}$ activation observed in the test set (10K samples) for the digits of each color in the last convolutional layer of the models trained in a biased way according to the three color distributions (a-c) and for the unbiased model (d). In brackets the classification accuracy obtained for the test set.

bias also affects the activations. On the one hand, a model trained with a color bias tends to produce higher activations for this color (see Fig. 3.a-c). On the other hand, a model trained without bias (i.e. a uniform distribution of colors in the training set) produces similar activations for the different colors (see Fig. 3.d). These results suggest the potential of the proposed $\lambda^{[l]}$ activation as a good estimator of the bias in a trained model.

2) Gender Recognition: Table I shows the results of the experiments described in Section IV-C, performed on *Networks* 1 and 2, respectively (see Section IV-B). The results show that the models trained using data from a single ethnic group perform better for this group (*Biased Models*). These results suggest that the ethnic features affect the performance of gender recognition based on the two popular network architectures evaluated. The learned parameters $\mathbf{w}$ of a model trained with one ethnic group (for example ethnic group $k = 1 \equiv$ *Asian*) do not generalize in the best possible way for other groups, with a clear drop in performance between testing groups for both networks, i.e., using the notation introduced in Section III-A: leaving fixed the network trained for $T =$ *Gender classification* by maximizing the goodness criterion G_T over $\mathcal{D}^{Asian}_{Ethnicity}$ (see Equation 1), we observe that $G_T(\mathcal{D}^{Asian}_{Ethnicity}) >> G_T(\mathcal{D}^{African}_{Ethnicity})$ and $G_T(\mathcal{D}^{Caucasian}_{Ethnicity})$.

On the other hand, training another network also for *Gender classification* in this case with unbiased data representing well all ethnic groups (*Unbiased Models*) and leaving it fixed, reduces the performance gap between testing groups and improves the overall accuracy (Avg in Table I) i.e. $G_{Gender}(\mathcal{D}^{Asian}_{Ethnicity}) \approx G_{Gender}(\mathcal{D}^{African}_{Ethnicity}) \approx G_{Gender}(\mathcal{D}^{Caucasian}_{Ethnicity})$. However, the performance achieved by the *Unbiased Models* trained with heterogeneous data does not improve the best performance achieved by each of the *Biased Models* trained using data only from one ethnic group.

B. InsideBias: Activation as a bias estimator

Fig. 4 shows the *Normalized Activation* from Equation 5 of the different networks for each demographic group. Fig. 4 shows that the activations obtained for the *Unbiased Models* in the last layers have the lowest differences between ethnic groups in testing. As we can see in the activation curves by layer, for *Biased Models* the differences in the activation between well-represented and poorly-represented groups are not homogeneous across layers. The curves suggest greater activation differences between groups, in this case (*Biased Models*), specially in the last layers of the networks.

TABLE I

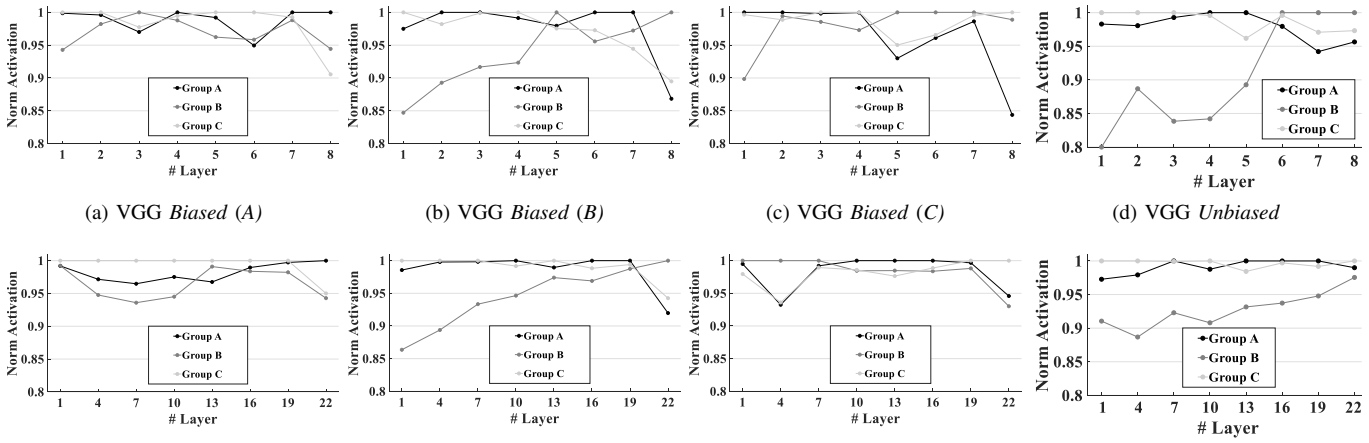
ACCURACY (%) IN GENDER CLASSIFICATION FOR VGG AND RESNET MODELS FOR EACH OF THE THREE DEMOGRAPHIC GROUPS. EACH LINE INDICATES A MODEL. THE PROTOCOL COLUMN INDICATES THE ETHNIC GROUP EMPLOYED TO TRAIN, THE GROUP COLUMNS INDICATE THE TESTING GROUPS, AND THE OTHER COLUMNS: AVERAGE ACCURACY ACROSS GROUPS (AVG) AND STANDARD DEVIATION (STD) (LOWER MEANS FAIRER)

VGG					
Protocol	**A**	**B**	**C**	**Avg**	**Std**
Biased (*A*)	**95.72**	94.16	94.68	94.85	0.65
Biased (*B*)	94.16	**95.82**	94.06	94.16	1.31
Biased (*C*)	92.46	94.63	**96.71**	94.60	1.74
Unbiased	94.84	95.69	95.28	95.27	0.34
ResNet					
Protocol	**A**	**B**	**C**	**Avg**	**Std**
Biased (*A*)	**96.84**	94.14	94.45	95.14	1.21
Biased (*B*)	93.29	**96.86**	95.40	95.18	1.47
Biased (*C*)	94.80	95.21	**97.01**	95.67	0.96
Unbiased	95.50	95.35	96.11	95.65	0.33

We also see in Fig. 4 in the first layers that testing **Group B** gets considerably less activation than the other two groups (mainly in the models trained with this group, see 4b and 4f), which tells us that this group has less activation for layers extracting low level features (e.g. shape, texture, and colors). However, this lower activation in the first layers does not necessarily imply a low performance (as seen in Table I). The low activation in the first layers is compensated with high activation in the last layers which are related with high level features close to the task T (i.e. gender recognition in our experiments). These results suggest a correlation between the bias introduced in Section IV-C, the performance reported in Table I, and the activations showed in Fig. 4.

C. InsideBias: Detecting bias with very few samples

The activations presented in Fig. 4 shows the relationship between bias and activations over 18K images, 2 learning architectures, and 5 models trained according to different biased



(a) VGG *Biased* (*A*) (b) VGG *Biased* (*B*) (c) VGG *Biased* (*C*) (d) VGG *Unbiased*

(e) ResNet *Biased* (*A*) (f) ResNet *Biased* (*B*) (g) ResNet *Biased* (*C*) (h) ResNet *Unbiased*

Fig. 4. Normalized activation λ' observed in **testing** (18K images) for the three demographic **Groups** (**A**, **B** and **C**) of the different *trained* models: *Biased Model* (*A*, *B* and *C*) and *Unbiased Model*. The top row shows the activations of *Network* 1 (VGG) and the bottom row plots the activations of *Network* 2 (ResNet). The VGG model has only 8 convolutional layers while the ResNet model has 22 (only the main ones appear).

and unbiased datasets. The following experiments investigate how InsideBias performs for bias detection using a small set of test samples.

Table II shows the average classification scores and the *Activation Ratio* defined in Equation 6 (l = last layer) obtained with the different considered models (Biased and Unbiased) trained for gender classification and tested with the 15 face images shown in Fig. 2. The results show how activations are correlated with biases even if the classification scores show almost no differences between biased and unbiased models. An *Activation Ratio* close to one reveals a similar activation pattern for all demographic groups tested, which is what we have for the *Unbiased models* (i.e. $\lambda^{[l]}(\mathcal{D}^{Asian}_{Ethnicity}) \approx \lambda^{[l]}(\mathcal{D}^{African}_{Ethnicity}) \approx \lambda^{[l]}(\mathcal{D}^{Caucasian}_{Ethnicity})$). In contrast, the *Biased models* show lower *Activation Ratio*, which means higher differences between activation patterns from well-represented and poorly-represented demographic groups (i.e. $\lambda^{[l]}(\mathcal{D}^{Well\text{-}represented}_{Ethnicity}) > \lambda^{[l]}(\mathcal{D}^{Poorly\text{-}represented}_{Ethnicity})$).

These results suggest that even if the network was trained only for gender recognition, the activation level of the filters is highly sensitive to the ethnic attributes. The proposed method for bias detection in deep networks, InsideBias, consists of measuring that sensitivity with the *Activation Ratio* $\Lambda_d^{[l]}$ defined in Equation 6 and comparing it to a threshold τ.

The main advantage of this method for the analysis of bias with respect to a performance-based evaluation is that the differences are examined in terms of model behavior. Images of Fig. 2 obtained good performance (over 99.99% confidence score even in biased models) but showed clearly different activation patterns λ. Bias analysis based on performance require large datasets, and using the proposed *Activation Ratio*, few images may be enough to detect biased models.

In this work we do not underestimate the performance as a good instrument for analyzing bias in deep networks. We propose to include activation as an additional evidence [38], which is specially useful when only very few samples are available for bias analysis.

TABLE II
AVERAGE GENDER CONFIDENCE SCORES S AND ACTIVATION RATIOS $\Lambda_d^{[l]}$ OBTAINED BY THE BIASED AN UNBIASED MODELS TESTED FOR THE 15 IMAGES OF FIG. 2. l = LAST CONVOLUTIONAL LAYER. 1 IS 100% CONFIDENCE ABOUT THE TRUE GENDER ATTRIBUTE IN THE IMAGE.

Model (*Training*)		**Test Group** A	B	C	$\Lambda_d^{[l]}$
VGG-Biased (*A*)	S	1.000	1.000	1.000	-
	$\lambda^{[l]}$	2.90	2.89	2.41	0.83
VGG-Biased (*B*)	S	1.000	1.000	1.000	-
	$\lambda^{[l]}$	2.24	3.25	2.61	0.69
VGG-Biased (*C*)	S	1.000	1.000	1.000	-
	$\lambda^{[l]}$	2.36	2.52	2.86	0.82
VGG-Unbiased	S	1.000	1.000	1.000	-
	$\lambda^{[l]}$	2.49	2.67	2.51	0.93
ResNet-Biased (*A*)	S	1.000	1.000	1.000	-
	$\lambda^{[l]}$	2.82	2.65	2.53	0.90
ResNet-Biased (*B*)	S	0.999	1.000	1.000	-
	$\lambda^{[l]}$	2.11	2.47	2.35	0.85
ResNet-Biased (*C*)	S	1.00	1.000	1.000	-
	$\lambda^{[l]}$	2.11	2.32	2.35	0.90
ResNet-Unbiased	S	0.999	0.999	0.999	-
	$\lambda^{[l]}$	2.33	2.32	2.34	0.99

VI. CONCLUSIONS

In this work we presented a preliminary analysis of how biased data affect the learning processes of deep neural net-

work architectures in terms of activation level. We showed how ethnic attributes affect the learning process of gender classifiers. We evaluated these differences in terms of filter activation, besides performance, and the results showed how the biases are encoded heavily in the last layers of the models. This activation reveals behaviors usually hidden during the learning process. We also evaluated different training strategies that suggest to what extent biases can be reduced if the whole network is trained using a heterogeneous dataset.

We finally propose a novel method, InsideBias, to detect bias through layer activations. InsideBias has two major advantages with respect to detection based on performance differences across demographic groups: 1) it does not require many samples (we showed that biased behaviors can be detected with only 15 images), and 2) InsideBias can give an indication of the bias in the model using only good samples correctly recognized (even with the highest confidence).

VII. Acknowledgments

This work has been supported by projects BIBECA (RTI2018-101248-B-I00 MINECO/FEDER), TRESPASS (MSCA-ITN-2019-860813), PRIMA (MSCA-ITN-2019-860315), and Accenture. I. Serna is supported by a research fellowship from the Spanish CAM.

References

[1] P. Stone, R. Brooks, E. Brynjolfsson, R. Calo, O. Etzioni, G. Hager, J. Hirschberg, S. Kalyanakrishnan, E. Kamar, S. Kraus *et al.*, "Artificial Intelligence and Life in 2030," *One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel*, p. 52, 2016.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. B. Estrach, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing Properties of Neural Networks," in *International Conference on Learning Representations (ICLR)*, Banff, Canada, 2014.

[3] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Demographic Effects in Facial Recognition and Their Dependence on Image Acquisition: An Evaluation of Eleven Commercial Systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, pp. 32–41, 2019.

[4] J. A. Buolamwini, "Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers," Ph.D. dissertation, Massachusetts Institute of Technology, 2017.

[5] M. Alvi, A. Zisserman, and C. Nellåker, "Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network embeddings," in *European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018.

[6] A. Acien, A. Morales, R. Vera-Rodriguez, I. Bartolome, and J. Fierrez, "Measuring the Gender and Ethnicity Bias in Deep Models for Face Recognition," in *Iberoamerican Congress on Pattern Recognition (IAPR)*. Madrid, Spain: Springer, 2018, pp. 584–593.

[7] S. Nagpal, M. Singh, R. Singh, M. Vatsa, and N. Ratha, "Deep Learning for Face Recognition: Pride or Prejudiced?" *arXiv:1904.01219*, 2019.

[8] P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O'Toole, "An Other-Race Effect for Face Recognition Algorithms," *ACM Transactions on Applied Perception*, vol. 8, no. 2, p. 14, 2011.

[9] A. J. O'Toole, P. J. Phillips, X. An, and J. Dunlop, "Demographic Effects on Estimates of Automatic Face Recognition Performance," *Image and Vision Computing*, vol. 30, no. 3, pp. 169–176, 2012.

[10] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, "Face Recognition Performance: Role of Demographic Information," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1789–1801, 2012.

[11] H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic Estimation from Face Images: Human vs. Machine Performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1148–1161, 2015.

[12] P. J. Grother, M. L. Ngan, and K. K. Hanaoka, *Ongoing Face Recognition Vendor Test (FRVT) Part 2: Identification*, ser. NIST Internal Report. U.S. Department of Commerce, National Institute of Standards and Technology, 2018.

[13] ——, *Ongoing Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects*, ser. NIST Internal Report. U.S. Department of Commerce, National Institute of Standards and Technology, 2019.

[14] J. J. Howard, Y. Sirotin, and A. Vemury, "The Effect of Broad and Specific Demographic Homogeneity on the Imposter Distributions and False Match Rates in Face Recognition Algorithm Performance," in *International Conference on Biometrics Theory, Applications and Systems (BTAS)*. Tampa, Florida, USA: IEEE, 2019.

[15] I. Hupont and C. Fernandez, "DemogPairs: Quantifying the Impact of Demographic Imbalance in Deep Face Recognition," in *International Conference on Automatic Face & Gesture Recognition (FG)*. Lille, France: IEEE, 2019.

[16] B. Lu, J.-C. Chen, C. D. Castillo, and R. Chellappa, "An Experimental Evaluation of Covariates Effects on Unconstrained Face Verification," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, pp. 42–55, 2019.

[17] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Demographic Bias in Biometrics: A Survey on an Emerging Challenge," *IEEE Transactions on Technology and Society*, vol. 1, no. 2, pp. 89–103, 2020.

[18] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *European Conference on Computer Vision (ECCV)*. Zurich, Switzerland: Springer, 2014, pp. 818–833.

[19] J. Buolamwini and T. Gebru, "Gender Shades: Intersec-

tional Accuracy Disparities in Commercial Gender Classification," in *Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81, New York, NY, USA, 23–24 Feb 2018, pp. 77–91.

[20] I. Serna, A. Morales, J. Fierrez, N. Cebrian, M. Obradovich, and I. Rahwan, "Algorithmic Discrimination: Formulation and Exploration in Deep Learning-based Face Biometrics," in *AAAI Workshop on Artificial Intelligence Safety (SafeAI)*, New York, NY, USA, 2020.

[21] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing Higher-Layer Features of a Deep Network," University of Montreal, Tech. Rep. 1341, Jun. 2009.

[22] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," in *International Conference on Learning Representations (ICLR) Workshop*, Banff, Canada, 2014.

[23] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding Neural Networks Through Deep Visualization," in *Intenational Conference on Machine Learning (ICML) Deep Learning Workshop*, Lille, France, 2015.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks Via Gradient-Based Localization," in *International Conference on Computer Vision (CVPR)*. Honolulu, Hawaii, USA: IEEE, 2017, pp. 618–626.

[25] A. Nguyen, J. Yosinski, and J. Clune, "Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned by Each Neuron in Deep Neural Networks," in *Intenational Conference on Machine Learning (ICML) Deep Learning Workshop*, New York, NY, USA, 2016.

[26] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev, "The Building Blocks of Interpretability," *Distill*, vol. 3, no. 3, 2018.

[27] C. Olah, A. Mordvintsev, and L. Schubert, "Feature Visualization," *Distill*, vol. 2, no. 11, 2017.

[28] J. Oramas, K. Wang, and T. Tuytelaars, "Visual Explanation by Interpretation: Improving Visual Feedback Capabilities of Deep Neural Networks," in *International Conference on Learning Representations (ICLR)*, New Orleans, Louisiana, USA, 2019.

[29] B. Zhou, Y. Sun, D. Bau, and A. Torralba, "Revisiting the Importance of Individual Units in CNNs via Ablation," *arXiv:1806.02891*, 2018.

[30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[31] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim, "Learning Not to Learn: Training Deep Neural Networks With Biased Data," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, Nevada, USA: IEEE, 2019, pp. 9012–9020.

[32] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86(11), pp. 2278–2324, 1998.

[33] A. Morales, J. Fierrez, and R. Vera-Rodriguez, "SensitiveNets: Learning Agnostic Representations with Application to Face Recognition," *arXiv:1902.00334*, 2019.

[34] R. Ranjan, S. Sankaranarayanan, A. Bansal, N. Bodla, J.-C. Chen, V. M. Patel, C. D. Castillo, and R. Chellappa, "Deep Learning for Understanding Faces: Machines May Be Just as Good, or Better, than Humans," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 66–83, 2018.

[35] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez, "Facial Soft Biometrics for Recognition in the Wild: Recent Works, Annotation and COTS Evaluation," *IEEE Trans. on Information Forensics and Security*, vol. 13, no. 8, pp. 2001–2014, August 2018.

[36] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep Face Recognition," in *British Machine Vision Conference (BMVC)*, Swansea, UK, 2015.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, 2016, pp. 770–778.

[38] J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho, "Multiple Classifiers in Biometrics. Part 2: Trends and Challenges," *Information Fusion*, vol. 44, pp. 103–112, November 2018.