Deep Learning for Face Recognition: Pride or Prejudiced?

Shruti Nagpal[1], Maneet Singh[1], Richa Singh[1], Mayank Vatsa[1], Nalini Ratha[2]

[1]IIIT-Delhi, India, [2]IBM TJ Watsin Research Center, USA

{shrutin, maneets, rsingh, mayank}@iiitd.ac.in, ratha@us.ibm.com

Abstract

Do very high accuracies of deep networks suggest pride of effective AI or are deep networks prejudiced? Do they suffer from in-group biases (own-race-bias and own-age-bias), and mimic the human behaviour? Is in-group specific information being encoded sub-consciously by the deep networks?

This research attempts to answer these questions and presents an in-depth analysis of 'bias' in deep learning based face recognition systems. This is the first work which decodes if and where bias is encoded for face recognition. Taking cues from cognitive studies, we inspect if deep networks are also affected by social in- and out-group effect. Networks are analyzed for own-race and own-age bias, both of which have been well established in human beings. The sub-conscious behavior of face recognition models is examined to understand if they encode race or age specific features for face recognition. Analysis is performed based on 36 experiments conducted on multiple datasets. Four deep learning networks either trained from scratch or pre-trained on over 10M images are used. Variations across class activation maps and feature visualizations provide novel insights into the functioning of deep learning systems, suggesting behavior similar to humans. It is our belief that a better understanding of state-of-the-art deep learning networks would enable researchers to address the given challenge of bias in AI, and develop fairer systems.

1. Introduction

Deep Learning based systems show extremely high accuracy in many computer vision application such as object detection and recognition, and image synthesis [20, 22, 29]. In particular, significantly better results have been shown in many face analytic tasks including face detection, age and gender prediction and face recognition leading many researchers to claim *super human* performance in face recognition. However, recently reported bias in face analytics shows the issues with the current state of the art deep learning face analytic systems [28].
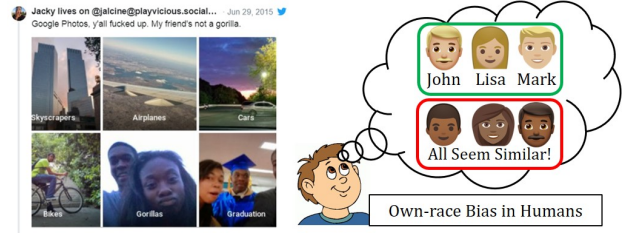
(a) (L-R): Recent incident of bias in AI (image taken from the internet); own-race bias already established in humans
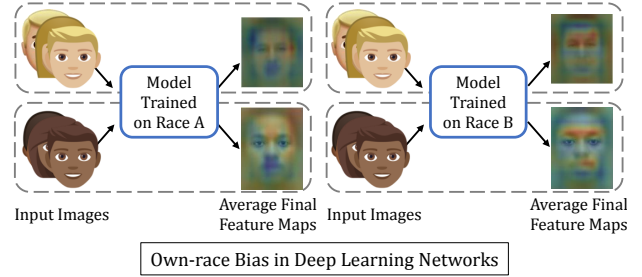
(b) Pride or Prejudiced? Own-race bias observed in deep learning models

Figure 1: Bias observed in (a) AI systems and humans, and (b) deep learning networks obtained via this research. Current state-of-the-art deep learning models appear to mimic the human in-group biases by focusing on different facial regions when trained on specific sub-groups for face recognition. Best viewed in color.

In July 2018, Amazon's facial recognition tool, 'Rekognition', falsely identified 28 members of the US Congress as criminals [1]. The US Congress contains around 20% members of color, out of which 40% were mis-classified. The study, conducted by the American Civil Liberties Union (ACLU), is one of the recent incidents shedding light on the emerging *biased* behavior of Artificial Intelligence systems, and the need to develop *fairer AI*. Earlier examples include the ProPublica study of algorithms being used by US courts and parole boards to predict future behaviour of criminals [3]. It was observed that the system was biased against black defendants, and gave them a higher score as compared to their white counterparts. Another popular instance which shook the Machine Learning community was

of the Google Photos application's mis-classification of a black couple as gorilla (Figure 1(a)) [2]. Multiple repeated incidents suggesting existence of bias demand dedicated research in order to ensure the development of AI systems for the *better*, and not for the *worse*. This research focuses on understanding the hidden biases of *face recognition systems*, specifically, current state-of-the-art deep learning networks. The existence of bias with respect to the race and age of an individual has been analyzed in face recognition models, by answering *if* and *where* the bias is encoded.

Bias is often defined as an *inclination or prejudice for or against one person or group, especially in a way considered to be unfair*. While bias in AI is a relatively new field, humans have shown to possess biases across different parameters since the very beginning. Neuro-cognitive researchers have long established the presence of *in-group* biases[1] in humans, wherein we tend to recognize individuals of our race or age easily, as compared to individuals of an *out-group* [27]. One such instance dates back to 1971, when five black men were acquitted for the murder of Khomas Revels [25]. Despite the lack of any forensic evidence, five eyewitnesses testified and identified them as offenders. However, during the third trial, private investigators located the actual culprits, who were later convicted. Interestingly, all the five men initially acquitted were black and all the five eyewitnesses were white. Social Psychologist, Dr. William Haythorn identified the reason for mis-identifications to be the cross-racial identifications, because of which people of the other race look alike, resulting in biased outcomes (Figure 1(a)). Several studies later, this phenomenon was termed as the *own race bias*, or *other race* or *cross-race effect*, where people of the other race look alike.

We believe that the human trait of making biased decisions has also made its way into the recent AI systems for computer vision (Figure 1(b)). Since most of these systems today are based on deep learning, they provide limited insights into the decision making process. Thus, understanding the models and their corresponding biases is of utmost importance, in order to prevent the inclusion of biases that existed in the past into our future.

1.1. Literature Review: Bias in Facial Analysis

A majority of research has focused on understanding bias in models developed on textual data [7], with limited attention to computer vision tasks [6]. However, recent research has focused on understanding the presence of bias in commercial-off-the-shelf systems and deep learning models primarily for gender and ethnicity classification in face images [8, 11, 32] and **not** face recognition. Buolamwini and Gebru [8] categorized face images into four categories based on their skin color: darker males, darker females,

lighter males, and lighter females. The performance of three commercial gender classification systems was analyzed with respect to the proposed categorization. It was observed that, out of the four categories, dark-skinned females were the most mis-classified group, thereby creating a need for fairer unbiased facial analysis algorithms. It is important to note that the authors analyzed gender classification using three commercial systems, without exploring the performance of deep learning models.

The establishment of bias for tasks such as gender classification, led to the requirement of *fairer* algorithms, capable of performing a particular task, unbiased of other related covariates. Das *et al.* [11] proposed a Multi-Task Convolutional Neural Network (MTCNN) for joint gender, race, and age classification of face images. The proposed model demonstrates improved performance for the given tasks across multiple subgroups of gender, age, and ethnicity. Simultaneously, Alvi *et al.* [4] proposed a joint learning and unlearning (JLU) technique for eliminating bias from CNN networks. Experiments are performed for multiple tasks of age, gender, race, and pose classification of face images, independently, while the other covariates are used for inducing bias. Ryu *et al.* [32] proposed using transfer learning to develop attribute prediction models inclusive of different race and gender subgroups. The proposed InclusiveFaceNet demonstrates improved performance of attribute classification across different subgroups.

1.2. Research Contributions

This research extends beyond the existing literature by analyzing deep learning based *face recognition models* for bias existence. This is the first of its kind research which attempts to understand what deep networks are encoding, and whether they learn features similar to the human brain, with respect to different co-variates of face recognition. Four deep learning networks, (LightCNN-9 [36], LightCNN-29 [36], ResNet50 [17], and SENet50 [19]), either trained from scratch or pre-trained on large scale datasets containing about 10M images have been evaluated to establish our findings. The key findings of this research are:

- **Result-1:** Upon simulating the real world scenario of limited exposure to other races/age-groups by training on data belonging to a specific class, we observe that deep learning networks mimic the human tendency of *in-group bias* across race and age.

- **Result-2:** Similar to the human behavior of face recognition, race-specific discriminative *regions of interest* are encoded in networks trained on a particular race.

- **Result-3:** Extending upon the existing cognitive literature, different regions of interest are also observed for varying age-groups, suggesting *own-age bias* in deep learning networks.

---

[1]In-group bias refers to the tendency to have an affinity towards members belonging to the same group as the individual.

- **Result-4:** Networks pre-trained on large-scale data with varying distribution (eg. VGGFace2 [9] and MSCeleb-1M [16]), demonstrate high generalization abilities, however, bias across age persists to be a major challenge.
- **Result-5:** Fine-tuning a pre-trained network results in change of *region of interest*, which might reduce its generalization abilities.

It is our belief that insights into the functioning of deep networks can help develop fairer AI systems, capable of unbiased decisions. Since deep learning models appear to mimic the human brain, findings of this research can also enable better utilization of pre-trained systems by applying suitable cognitive techniques (designed for humans) for de-biasing. The remainder of this paper is organized as follows: the following section provides the experiments and protocols used for both the case studies of race and age. Sections 3 and 4 present the findings, along with the key takeaways and insights, followed by the conclusions of this research.

2. Experiments and Protocols

In this research, deep-learning based face recognition networks are analyzed to answer two key questions:

- Does deep learning encode race-specific information?
- Does deep learning encode age-specific information?

We analyze the behavior of deep learning networks with respect to well established cognitive studies on humans, and identify the regions used by deep learning models to learn discriminative features for face recognition. We further delve deeper to observe if these regions are consistent across different races and age-groups, and how deep learning models behave when trained from scratch on selective data i.e. data highly biased towards specific subgroups.

Network Details: Four deep learning based face recognition networks are analyzed to answer the two key questions pertaining to the effect and existence of bias due to race and age: (i) LightCNN-9 [36], (ii) Light CNN-29 [36], (iii) ResNet50 [17], and (iv) SENet50 [19]. LightCNN-9 is trained from scratch, while the remaining three networks are pre-trained on *large-scale* face datasets. LightCNN-29 is pre-trained on the MS-Celeb-1M dataset[2] [16], which contains 10M images corresponding to nearly 10k identities in the training set. ResNet50 and SENet50 are pre-trained on the VGG-Face2 [9] and MS-Celeb-1M datasets for face recognition[3]. VGG-Face 2 dataset contains 3.31M images pertaining to 9,131 identities. Therefore, the two models are trained on over 13M images, the largest corpus used for training a single face recognition model.

Datasets Used for Case-study-1 (Effect of Race): The behavior of deep learning networks is analyzed with respect

Figure 2: Samples from the datasets used for understanding racial bias in face recognition: (i) CMU Multi-PIE (Caucasoid race), (ii) MORPH (Negroid race), and (iii) RFW: Caucasian (Caucasoid race) and African (Negroid race).
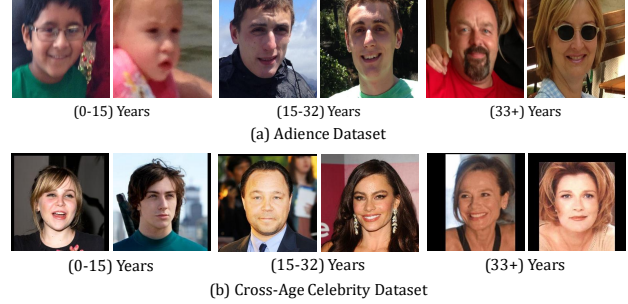

(0-15) Years   (15-32) Years   (33+) Years
(a) Adience Dataset

(0-15) Years   (15-32) Years   (33+) Years
(b) Cross-Age Celebrity Dataset

Figure 3: Sample images of the two datasets used for analyzing the effect of varying age on deep learning networks.

to two races: *Caucasoid (white)* and *Negroid (black)*. Three datasets are used for performing the said analysis:

(i) CMU Multi-PIE [15]: Over 44K images of 336 subjects images are selected which correspond to frontal face images having illumination and expression variations.

(ii) Craniofacial Longitudinal Morphological (MORPH) Album-2 [30]: Over 52k frontal face images pertaining to 10,409 subjects are used.

(iii) Racial Faces in the Wild (RFW) [34] is an unconstrained dataset of African, Caucasian, Asian, and Indian face images, collected from the Internet. Its test set contains labeled images *across races*, with limited labeled Caucasian faces in the training set. For analysis, the test set containing 10,145 Negroid and 10,196 Caucasoid face images of 2,995 and 2,959 subjects, respectively, has been used.

The CMU Multi-PIE dataset is used for analyzing the networks for the Caucasoid race, while the MORPH dataset is used for the Negroid race. Both the datasets are captured in constrained settings, and for both, 70% of the subjects are used to create the training (or fine-tuning) partition, while the remaining 30% subjects form the test set. This ensures disjoint training and testing partitions, in terms of images and subjects. Due to the availability of limited labeled training data in the RFW datset, only the test data is used to evaluate our hypothesis. Therefore, unless explicitly specified, analysis is drawn using the CMU Multi-PIE and MORPH datasets. Figure 2 presents sample images from the three datasets demonstrating facial images across the two races.

Datasets Used for Case-study-2 (Effect of Age): Deep learning networks are analyzed for face recognition with

respect to three age groups: (i) (0-14) years, (ii) (15-32) years, and (iii) (33+) years. To the best of our knowledge, there does not exist any large scale dataset containing the identity and age information for such a large range. IMDB-Wiki [31] is a large dataset containing face images with age and identity information, however, it has been observed that the dataset contains erroneous age information. Therefore, in this study we combined face images from two datasets: Adience dataset [12] and Cross-Age Celebrity Dataset (CACD) [10]. Both the datasets contain images collected from the Internet, having pose, illumination, and background variations. Figure 3 presents sample images from both the datasets, displaying variations in age. The CACD dataset contains over 1,60,000 face images belonging to the age range of 14-62 years. In our experiments, the CACD dataset has been used for training (or fine-tuning) the deep learning networks. However, since it does not contain subjects below the age of 14 years, 70% of the Adience dataset has also been used for training. Thus, the training or fine-tuning set is a combination of images from CACD and Adience datasets. The remaining 30% of data pertaining to the Adience dataset is used for testing. Mutual exclusion (for both images and subjects) has been ensured between the training and testing partitions.

Details Regarding Experiments and Analysis: Analysis is drawn from 36 experiments (Table 1) conducted across different case-studies, networks, and training data. At the time of training from scratch, models are trained on data pertaining to a single sub-group only (race or age), to observe the discriminative regions being learned for face recognition; and on equal number of samples belonging to different subgroups. Cross sub-group experiments have also been performed on the trained models to understand what regions are being used in each case to perform classification. For example, LightCNN-9 trained on the Caucasoid subjects is used for evaluation on the Negroid subjects, to study the difference in processing (if any). Three pre-trained networks ResNet50, SENet50 and LightCNN-29 have also been evaluated for performance and visualizations. The effect of fine-tuning has also been studied to understand how the discriminative regions evolve for different sub-groups with pre-trained networks. Fine-tuning is performed on the entire network for the given train set. Feature-level analysis is performed with respect to the learned feature maps, where the final convolution layer maps are interpolated and super-imposed on the input image, to obtain the region of interest for the given input image and network. Class Activation Maps (CAMs) [37] are used to obtain the most discriminative regions of face images, focused upon by the CNNs. Results are also reported in terms of the verification accuracy (Genuine Acceptance Rate) obtained at 1% False Acceptance Rate. In order to ensure consistency across analysis and fair comparisons, fixed

Table 1: List of 36 experiments performed to understand face recognition in deep learning models w.r.t. race and age bias. Results have been shown using three datasets for race: CMU Multi-PIE, MORPH, and RFW; and two datasets for age: Adience and CACD. LCNN: LightCNN; MSC: MS-Celeb-1M, VF2: VGG-Face2; C: Caucasoid, N: Negroid; A1: (0-14) years, A2: (15-32) years, A3: (33+) years.

Exp.#	Network	Pre-train	Train	Fine-tune	Test
Effect of Race					
1 / 2	LCNN-9	-	C	-	C / N
3 / 4			N	-	C / N
5 /6			C, N	-	C / N
7 / 8	LCNN-29	MSC (10M)	-	-	C / N
9-12				C / N	C / N
13 / 14	ResNet50	MSC (10M),	-	-	C / N
15-18			-	C / N	C / N
19 / 20	SENet50	VF2 (3M)	-	-	C / N
21-24			-	C / N	C / N
Effect of Age					
25	LCNN-9	-	A1	-	A1
26			A2	-	A2
27			A3	-	A3
28-30	LCNN-29	MSC (10M)	-	-	A1/A2/A3
31-33	ResNet50	MSC (10M),	-	-	A1/A2/A3
34-36	SENet50	VF2 (3M)	-	-	A1/A2/A3

protocols have been used for all experiments. Model files will be released for reproducibility of findings.

3. Does Deep Learning Encode Race-specific Information?

The *own-race bias* in humans has been well established, where we are able to recognize people of our own race more easily as compared to individuals of other race [26, 14]. Also referred to as the *other-race bias* or *cross-race effect*, it has extensively been studied in humans in the form of cognitive, behavioural, and neuro-imaging studies. Depending on the racial identity of the individual and the person to be identified, different facial regions have shown more contribution during the decision making process. In this study, we understand the behaviour of deep learning models to observe if they also focus on specific discriminative regions based on the race, as observed in humans.

3.1. Are Deep Learning Networks Prejudiced?

Experiments 1-6 (Table 1) are performed to analyze the presence of prejudice in deep learning networks via class activation maps and feature visualizations. LightCNN-9 models are trained from scratch on data belonging to different races. The aim is to understand if faces of different races are learned in a similar manner, or does the model utilize race-specific features for identifying individuals.

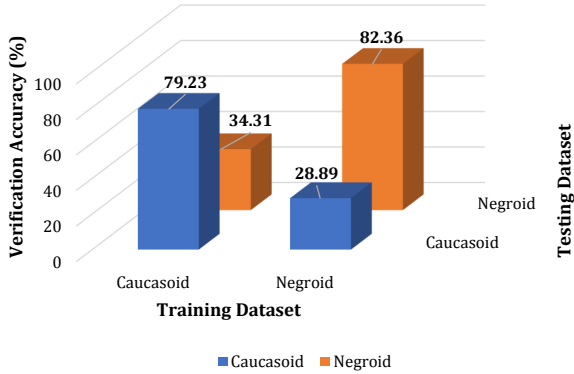Figure 4 presents the GAR values obtained at 1%FAR with LightCNN-9 models, for the two races. The model

Figure 4: Verification accuracy (%) at 1% False Acceptance Rate of LightCNN-9 architecture. Models trained on images pertaining to a particular race demonstrate poor performance on face images of the other race.

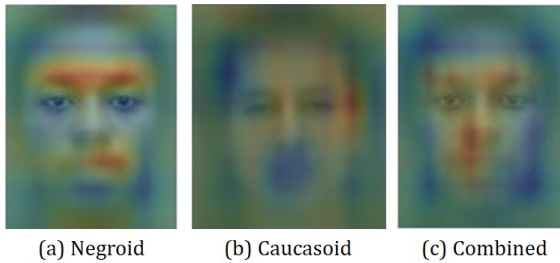(a) Negroid   (b) Caucasoid   (c) Combined

Figure 5: Visualization of salient regions obtained from exp. 1-5 (Table 1), where networks are trained from scratch on (a) only Negroid, (b) only Caucasoid, and (c) both Negroid and Caucasoid face images. Best viewed in color.

trained on face images of Caucasoid faces only reports an accuracy of 79.2% on the testing set of same race (Caucasoid). On the other hand, the model trained on Negroid faces, tested on the Caucasoid individuals, reports a GAR of 28.9%. This suggests the presence of race-specific prejudice being encoded in the network due to the variations in the input data. Similar trends are observed when testing race-specific models on data belonging to Negroid individuals only. The LightCNN-9 network trained on faces of Negroid subjects only reports a GAR of 84.3% on samples of the Negroid race (same race), while the network trained on Caucasoid faces achieves a GAR of 34.3% (different race).

We also evaluated the models trained on a specific race on the RFW test set, and observed a similar trend, thus demonstrating that the drop in accuracy is caused due to the other-race bias and not dataset bias. When testing on the RFW Caucasian test set, the network trained on Negroid race in comparison to the network trained on the Caucasoid race, demonstrates a decrease of around 22% of its performance. It is our belief that network prejudice occurs due to differences in the learned networks for varying input data.

Feature visualizations allow us to compare models beyond the verification accuracy, and analyze the regions be-
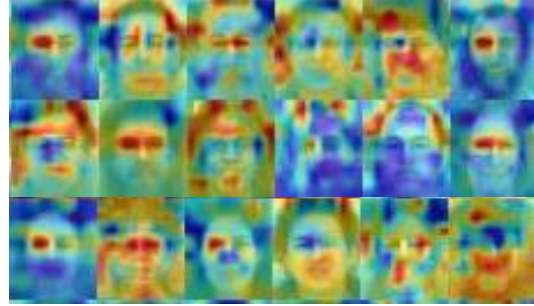

Figure 6: Sample average class activation maps obtained from the trained the LightCNN-9 model trained on Caucasoid subjects. The model focuses heavily on the eye region for recognition. Best viewed in color.

ing learned by the network to extract features with the highest discriminative information across races. Figure 5(a-b) presents the salient regions used for feature extraction by the learned networks. These are obtained by interpolating the final convolution layer filter responses and superimposing on the input image. It can be observed that both models focus on different regions for feature extraction. Figure 5(c) shows the visualizations of salient regions learned by a LCNN-9 model trained on equal Caucasoid and Negroid faces. The network learns a union of salient regions obtained for each race independently. While for the Negroid race, salient regions correspond to the lips and above the eyes, Caucasoids appear to be distinguished on a more holistic view, with more focus on face boundary. *The differences obtained in the feature maps, along with the accuracy variations strengthen the hypothesis that faces belonging to different race are encoded differently in a classification model, thereby suggesting presence of bias.*

3.2. Do Deep Networks Mimic Humans?

Cognitive studies in humans have analyzed how Negroid and Caucasoid individuals perform face recognition [13, 18]. Across different studies, it has been observed that Negroid participants focus more on certain facial features such as mouth, lips, and nose while identifying other individuals of the same race [13]. On the other hand, Caucasoid subjects used traits such as the iris color, face shape, hair color and texture for describing and identifying other people of the same race [13, 33]. Such studies suggest a higher level of difference between the regions useful for distinguishing between individuals of different race. For example, information such as the eye color or hairstyle might not be useful for identifying a Negroid individual, whereas it could be of utmost importance for identifying a Caucasoid.

Intrigued by the findings of human behavior, in this study, deep learning based face recognition systems are analyzed to investigate if they follow a similar pattern. In order to understand the behavior of deep learning networks, in terms of the discriminative regions of interest, class ac-

(i) Pre-trained SENet50 | (ii) Pre-trained ResNet50 | (iii) Pre-trained LightCNN-29 | (vi) Trained LightCNN-9 | (v) Fine-tuned SENet50 | (vi) Fine-tuned ResNet50

(a) Negroid

(i) Pre-trained SENet50 | (ii) Pre-trained ResNet 50 | (iii) Pre-trained LightCNN-29 | (vi) Trained LightCNN-9 | (v) Fine-tuned SENet50 | (vi) Fine-tuned ResNet50

(b) Caucasoid

(I) Feature Visualizations

(a) Original Images | (b) SENet50 Features

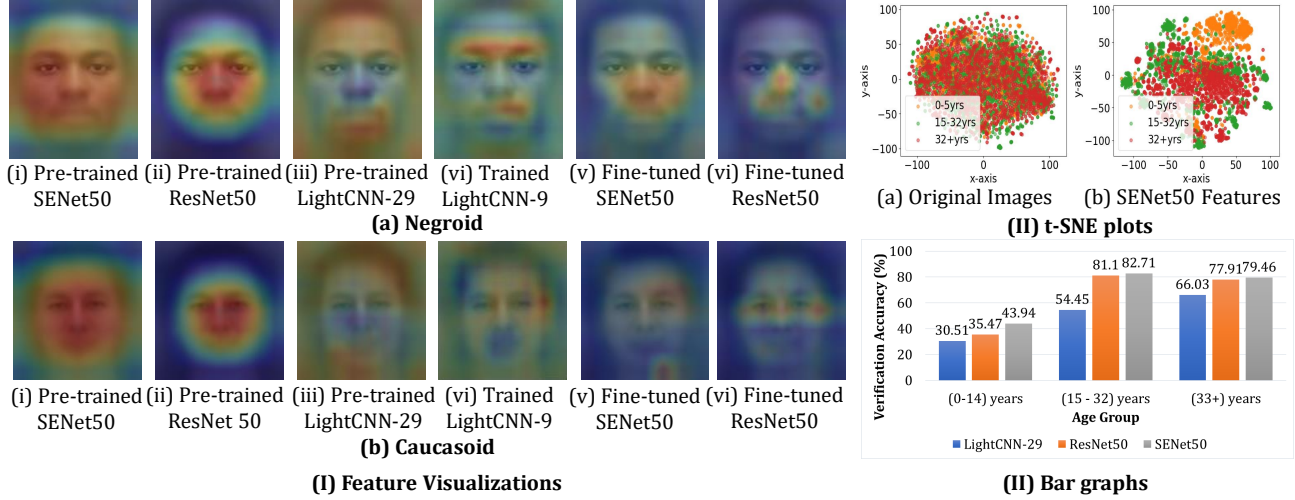**(II) t-SNE plots**

(II) Bar graphs

Figure 7: (I) Visualization of salient regions obtained from pre-trained, fine-tuned, and trained from scratch networks. Differences in the *region of focus* are observed across models. (II) t-SNE visualizations of raw pixels and features extracted from SENet50. (III) Verification accuracy (%) at 1%FAR for three pre-trained networks on the three age based sub-groups.

Figure 8: Sample average class activation maps obtained upon from the LightCNN-9 model on trained Negroid subjects. The model focuses heavily on the nose region (below the eyes), and the chin region. Best viewed in color.

tivation maps of the LightCNN-9 models trained on data pertaining to Caucasoid and Negroid races for face recognition are analyzed (Experiment 1-6, Table 1). Networks trained on a particular race simulate the functioning of the human brain which has been in contact with individuals of a specific race only, thereby enabling us to understand own-race effect in deep networks. Figures 6 and 8 present sample mean class activation maps. Each map corresponds to the mean activation associated with a particular class. It is interesting to note that the activation maps vary significantly between races, however, demonstrate a similar behavior within a particular race. The network trained on subjects belonging to Caucasoid race only, discriminates primarily on the basis of the eye region (Figure 6). These results are in conjunction with the cognitive studies reported in literature, where researchers have identified the eye color as one of the most discriminative traits used by humans as well [13]. *This suggests that the behavior of deep learn-*

ing networks and the human brain is alike. Apart from the eyes, we also observe regions around the face being highlighted, which could mean face shape or hair type, both of which have been identified earlier from behavioural studies [13]. Similarly, Figure 8 presents sample mean class activation maps from the LightCNN-9 network trained on the Negroid dataset. Regions of the lip, nose, and cheekbones appear to contribute more towards face recognition, as compared to other facial regions. Similar results have also been reported in cognitive studies, wherein Negroid individuals utilized the eyes and lips for describing and recognizing other Negroid subjects [13]. *The regions of interest used by models trained on specific races demonstrate the presence of different discriminative regions, suggesting that networks sub-consciously encode race-specific information, thereby resulting in an own-race bias.*

3.3. More the Merrier: Does Large-scale Data Help in Mitigating Own-race Bias?

The presence of own-race bias in individuals is often attributed to limited exposure to other race people [23]. In literature, studies have shown reduced effect of other-race bias upon training subjects to recognize face images of a different race. Since deep learning models rely heavily on the training samples, networks trained on large datasets may exhibit a similar behavior. In order to understand whether pre-trained networks suffer from the challenge of bias, deep networks trained on large-scale datasets are analyzed (Table 1: Exp. 7-24). The performance of three pre-trained networks - ResNet50, SENet50, and LightCNN-29 is studied. Feature maps and class activation maps are also computed to better understand their behaviour. Upon testing on the relatively constrained data belonging to the Caucasoid race, ResNet50 obtains an accuracy of 98.7%, while for the

Negroid race, the network attains a classification accuracy of 96.3%. Similar results are obtained with the SENet50 network: 98.8% and 96.5%, respectively. The performance of LightCNN-29 network further depicts that the generalization of a network highly depends upon the training data used. It achieves an accuracy of 96.47% and 78.12% on the Caucasoid and Negroid data, respectively. The results depict that pre-trained models are biased towards the Caucasoid race and achieve better performance on Caucasoid faces as compared to samples belonging to the Negroid race.

Similar trend in results is observed on the unconstrained in-the-wild RFW dataset. The pre-trained ResNet-50 network achieves 77.3% on the Negroid race and 90.6% on the Caucasoid race, while the pre-trained SENet-50 model attains an accuracy of 76.4% on the Negroid race versus 90.4% on the Caucasoid race. Similar to the performance on the constrained data, the performance of the LCNN-29 model reduces to 58.2% and 82.1% for the Negroid and Caucasoid race, respectively. The consistent reduced performance on the Negroid race suggests that the generalization capability of a network depends heavily upon the amount and variability of the training data. Figure 7 presents the feature visualizations (region of interest) for the pre-trained networks. Both ResNet50 and SENet50 models learn feature maps which cover an entire circular region of the face, thereby using almost the entire face for recognition, which might result in higher generalizability.

Fine-tuning has emerged as a common practice for enhancing a network's performance for a chosen task and dataset. Experiments 9-12, 15-18, and 21-24 of Table 1 are performed to understand the behaviour of deep networks after fine-tuning. It is interesting to note the transition in *region of interest* from pre-trained networks to fine-tuned networks (Figure 7). Initially, for both SENet50 and ResNet50, feature maps of both the races (Caucasoid and Negroid) are similar, and focus on a circular region covering almost the entire face. After fine-tuning, the updated feature maps appear similar to the ones discussed earlier (Figure 5). The feature maps obtained after fine-tuning with images of the Negroid race are similar to the network trained only on the Negroid race, demonstrating a shift to only the nose region and face center. In case of Caucasoid individuals, the networks start focusing on the eye-region only. This demonstrates the importance of fine-tuning for specific problems. However, it also establishes that *fine-tuning should only be performed when the test set is highly specific as it might reduce the generalization abilities of the network*.

4. Does Deep Learning Encode Age-specific Information?

Similar to the own-race bias, recently researchers have established the presence of *own-age bias* in human beings [5]. Own-age bias refers to the phenomenon of an individual being able to identify other people of a similar/same age with a greater accuracy (and more easily) as compared to individuals of other ages. Research has focused on understanding the own-age bias on three age-groups: children, adults, and older people. While initial studies appeared conflicted in terms of an own-age bias in older people, however, across different studies, researchers have established its presence across different age groups [35]. This is the first research which analyzes and studies the bias effect due to the factor of age in deep learning models. Previously, Kemelmacher-Shlizerman *et. al* [21] presented the MegaFace dataset and studied the impact of age on face recognition with respect to the data used for training a network. However, no analysis of deep models has been performed for the same. This section focuses on analyzing the own-age bias in deep learning models, specifically in terms of questioning its existence and investigating if distinguishing features vary across age.

4.1. How Well Do Deep Networks Recognize Children, Youngsters, and Adults?

In order to analyze the behavior of deep learning networks in terms of own-age bias, a similar set of experiments are performed as the previous case study. Models (both trained from scratch (Table 1 1: exp. 25-27) and pre-trained (Table 1: expt. 28-26)) are analyzed on face images of varying age-groups. The pre-trained networks of ResNet50, SENet50 and LightCNN-29 are evaluated for the three different age-groups: (i) 0-14 years, (ii) 15-32 years, (iii) 33+ years. The networks which are pre-trained on a large set of data capturing multiple variations, that is, ResNet50 and SENet50 obtained an accuracy of 77.9% and 75.5%, respectively for the oldest age group of (33+) years. Both the networks achieve a slightly higher performance on the age-group of (15-32) years, by reporting around 81-83% (Figure 7(III)), and less than 45% for the youngest age group. On the other hand, LightCNN-29 reports a verification accuracy of 30.5%, 54.45%, and 66.03% for the three age groups, in ascending order of age. *The consistent lower face verification performance across different networks suggests the presence of an own-age bias in these networks.* Figure 7(II) presents the t-SNE [24] plots for the images of the test set (across all three age-groups), and the features extracted from SENet50. It is interesting to note that features of the youngest age-group (orange colour) appear to form a separate cluster, thereby suggesting that the network is able to distinguish between adults and children, however, the lower recognition performance suggest lower discrimination between the extracted features. The relatively improved performance of SENet50 and ResNet50 models (pre-trained on MS-Celeb-1M and VGGFace2) suggest that similar to the *contact hypothesis* given in cognitive studies, where more exposure to a particular out-group results

(0-15) Years   (15-32) Years   (33+) Years

Figure 9: Class activation maps of sample images obtained from the trained LightCNN-9 networks. Best viewed in color.

in increased recognition capabilities, more contact (training) with a sub-group of individuals might result in an improved classification performance for that sub-group. Thus, reinstating the importance of training data for modeling the different co-variates of face recognition.

4.2. What Does the Model See Based on the Age?

Cognitive research has established the presence of own-age bias, however, limited research has focused on analyzing the specific regions of focus or discriminative facial features across different age groups. While there exist studies which analyze the gaze pattern or scan pattern of individuals belonging to different age-groups, they do not contribute much to the current study, since they do not specifically analyze the discriminative regions of interest. In this study, we analyze the behavior of face recognition models to investigate different regions of interest.

Three LightCNN-9 models are trained, one for each age group, and corresponding class activation maps are presented in Figure 9. The class activation maps demonstrate the discriminative regions of focus used by the model for classifying the input images. An interesting trend of activation is observed across the images of different age groups. *Face images of kids (first four columns) appear to use more soft biometric information such as hair and face shape for classification, while models trained on individuals of (15-32) years focus more on the lip, lower face, and forehead regions. Models trained on the third category, (33+) years rely heavily on the eye information for classification, along with other facial features.* The activation suggests the presence of different discriminative regions across age-groups, thereby suggesting the presence of an own-age bias, especially for young children below 15 years. The presence of different discriminative regions across varying age-groups provides new insights into the face recognition process, which opens the avenues for novel research directions in both deep learning and cognitive studies.

5. Conclusion

Multiple instances have unfolded the *hidden biases* of AI systems, creating a need for developing *fairer systems*. In this direction, this research analyzes deep learning networks for the existence of bias in automated face recognition systems. This is the first work which attempts to
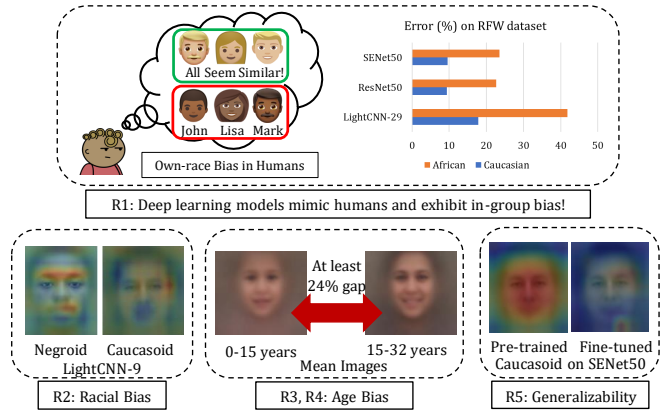

Figure 10: Deep learning for face recognition- Prejudiced!

answer *if* and *where* bias is encoded in face recognition systems, with respect to two case studies of race and age. Analysis across multiple deep learning networks suggests the presence of *own-race* and *own-age* bias in face recognition models. Similar to human behavior, deep learning networks demonstrate strong tendency of focusing on selected facial regions for a particular race, with variations across different race. To the best of our knowledge, this is the first study which suggests that the biased behavior of deep learning networks is similar to that observed in humans, therefore presenting opportunities of applying well established human cognitive results for bias elimination in deep learning networks. Our analysis also demonstrates the benefit of using large-scale data for training; however, does not recommend it as a complete solution for bias elimination. On the other hand, caution must be ensured while curating the training data for deep learning models, in order to incorporate maximum variability possible. Similar behavior is observed for the cross-age study, where face recognition networks appear to *sub-consciously* encode the age information pertaining to the training samples resulting in varying regions of interest across groups, and presence of *own-age bias*. Lower recognition performance of state-of-the-art face recognition systems on child face images further reaffirms the biased nature of such systems with respect to age as well. We believe that the insights provided in this paper can help facilitate research across multiple disciplines, and enable development of novel techniques for de-biasing existing face recognition systems.

References

[1] Facial recognition, and bias. https://tinyurl.com/y7rat8vb. 1

[2] Google photos labels dark skinned people as gorillas. https://tinyurl.com/y7c7dxso. 2

[3] Propublica study of algorithms. https://tinyurl.com/gvtccpq. 1

[4] M. Alvi, A. Zisserman, and C. Nellker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *European Conference of Computer Vision Workshops*, 2018. 2

[5] J. S. Anastasi and M. G. Rhodes. An own-age bias in face recognition for children and older adults. *Psychonomic bulletin & review*, 12(6):1043–1047, 2005. 7

[6] L. Anne Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 771–787, 2018. 2

[7] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *International Conference on Neural Information Processing Systems*, pages 4356–4364, 2016. 2

[8] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, 2018. 2

[9] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 67–74, 2018. 3

[10] B. Chen, C. Chen, and W. H. Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Transactions on Multimedia*, 17(6):804–815, 2015. 4

[11] A. Das, A. Dantcheva, and F. Bremond. Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In *European Conference of Computer Vision Workshops*, 2018. 2

[12] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, 2014. 4

[13] H. D. Ellis, J. B. Deregowski, and J. W. Shepherd. Descriptions of white and black faces by white and black subjects. *International Journal of Psychology*, 10(2):119–123, 1975. 5, 6

[14] S. Fu, H. He, and Z.-G. Hou. Learning race from face: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(12):2483–2509, 2014. 4

[15] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 3

[16] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102, 2016. 3

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2, 3

[18] P. J. Hills and M. B. Lewis. Reducing the own-race bias in face recognition by shifting attention. *The Quarterly Journal of Experimental Psychology*, 59(6):996–1002, 2006. 5

[19] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 3

[20] P. Hu and D. Ramanan. Finding tiny faces. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[21] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016. 7

[22] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua. Labeled faces in the wild: A survey. In *Advances in face detection and facial image analysis*, pages 189–248. 2016. 1

[23] S. Lebrecht, L. J. Pierce, M. J. Tarr, and J. W. Tanaka. Perceptual other-race training reduces implicit racial bias. *PloS one*, 4(1):e4215, 2009. 6

[24] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9:2579–2605, 2008. 7

[25] C. Meissner and J. Brigham. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7:3–35, 2001. 2

[26] C. A. Meissner and J. C. Brigham. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7(1), 2001. 4

[27] M. M. Nicholls, O. Churches, and T. Loetscher. Perception of an ambiguous figure is affected by own-age social biases. *Scientific Reports*, 8, 12 2018. 2

[28] I. D. Raji and J. Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *AAAI/ACM Conf. on AI Ethics and Society*, 2019. 1

[29] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135, 2019. 1

[30] A. W. Rawls and K. Ricanek. MORPH: Development and optimization of a longitudinal age progression database. In J. Fierrez, J. Ortega-Garcia, A. Esposito, A. Drygajlo, and M. Faundez-Zanuy, editors, *Biometric ID Management and Multimodal Communication*, pages 17–24, 2009. 3

[31] R. Rothe, R. Timofte, and L. V. Gool. DEX: Deep expectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops*, 2015. 4

[32] H. J. Ryu, H. Adam, and M. Mitchell. Inclusivefacenet: Improving face attribute detection with race and gender diversity. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2018. 2

[33] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948–1962, 2006. 5

[34] M. Wang, W. Deng, J. Hu, J. Peng, X. Tao, and Y. Huang. Racial faces in-the-wild: Reducing racial bias by deep unsupervised domain adaptation. *CoRR*, abs/1812.00194, 2018. 3

[35] H. Wiese, J. Komes, and S. R. Schweinberger. Ageing faces in ageing minds: a review on the own-age bias in face recognition. *Visual Cognition*, 21(9-10):1337–1363, 2013. 7

[36] X. Wu, R. He, Z. Sun, and T. Tan. A light CNN for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018. 2, 3

[37] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 4