Measuring the Gender and Ethnicity Bias in Deep Models for Face Recognition

Alejandro Acien, Aythami Morales, Ruben Vera-Rodriguez, Ivan Bartolome and Julian Fierrez

Biometrics and Data Pattern Analytics (BiDA) Lab, EPS,
Universidad Autonoma de Madrid C/ Francisco Tomas y Valiente 11, 28049 Madrid, Spain
{alejandro.acien, aythami.morales, ruben.vera, bartolome.gonzalez, julian.fierrez}@uam.es

Abstract. We explore the importance of gender and ethnic attributes in the decision-making of face recognition technologies. Our work is in part motivated by the new European regulation for personal data protection, which forces data controllers to avoid discriminative hazards while managing sensitive data like biometric data. The experiments in this paper are aimed to study what extent sensitive data like gender or ethnic origin attributes are present in the most common face recognition networks. For this, our experiments include two popular pre-trained networks: VGGFace and Resnet50. Both pre-trained models are able to classify gender and ethnicity easily (over 95% of performance) even suppressing 80% of the neurons in their embedding layers. The experimentation is conducted on a publicly available database known as Labeled Faces in the Wild with more than 13000 images of faces with a huge range of poses, ages, races and nationalities.

Keywords: Face recognition, human attributes, gender, ethnic, discrimination.

1 Introduction

Face recognitions systems have become popular due to good performance in human recognition, which has led this technology to take on a leading role in the last years. For example, common devices such as smartphones or laptops are applying face recognition for authentication and verification improving traditional recognition technologies based on passwords or swipe patterns. Advanced video surveillance also apply face recognition for continuous monitoring or intrusion detection with good results [1,2]. Additionally, various applications of face recognition are beyond identity management. The capacity to collect personal data has given advertisers the possibility to individualize marketing campaigns [3]. User profiling based on face images is a technology that brings the opportunity to collect gender, age and ethnicity from a picture of the face.

The performance of face recognition technology has been boosted by deep convolutional neuronal networks that have drastically reduced the error rates in the last decade [4]. These developments have resulted in a large variety of networks available for the research community and industry. Among the most popular stand out VGGFace [5], Resnet50 [6] and FaceNet [7]. These neuronal networks architectures are able to identify people with a face image with more than 97% of accuracy in the public dataset

Labeled Face in the Wild [8], furthermore the pre-trained models (both architecture and weights) of these networks are totally available and the use of them by research groups and commercial applications is growing continuously. As example of the impact of these pre-trained models, VGGFace, Resnet50 and FaceNet references [5,6,7] have achieved more than 12900 citations during the last 3 years according to Google Scholar.

A face image reveals information not only about who we are but also about what we are. During last decade, researchers have proposed to exploit auxiliary data of the users to improve face recognition [9,10]. Most of these auxiliary data, such as gender, ethnicity, age and behavior among others, can be easily inferred from a face picture. These auxiliary data are known as soft biometrics, which refer to those biometrics that can distinguish different groups of people but do not provide enough information to uniquely identify a person. Those attributes can be extracted with high accuracy (over 95%) using just one face picture [11].

Biometric technology and privacy of users have been confronting each other for a long time [12]. There is a never-ending trade-off between security of citizens and their privacy. Citizens and governments around the world are very conscious about data protection and personal information on the Big Data era. As a prove of this, in April 2016 the European Parliament adopted a set of laws aimed to regularize the collection, storage and use of personal information, the General Data Protection Regulation (GDPR) [13]. Biometric data is defined as sensitive data in this new GDPR due to its capacity to "*uniquely identifying a natural person*". This regulation is a step forward with respect to previous national and European laws and establishes the foundations of what anyone can do with data in this new era. Is biometric technology complying with this new regulation? According to paragraph 71 of GDPR, data controllers who process sensitive data have to "*implement appropriate technical and organizational measures*…" that "…*prevent, inter alia, discriminatory effects*". The discrimination is the unjust or prejudicial treatment of different categories of people, especially on the grounds of race, age, religion or gender. According to this definition, face images belong to this group of sensitive data regulated by the GDPR [14]. Facial attributes revealing the gender, age or ethnic have the potential to discriminate citizens based on the group to which that person belongs. It is important to note that we do not argue that face technology is discriminatory but rather, the hazard exists in case of unethical usages.

The aim of this paper is to study to what extent discriminative attributes such as gender or race can be obtained from feature vectors generated by state-of-the-art face recognition algorithms. This information is part of the decision making of any application based on these algorithms even if we use high-dimensional feature spaces trained for a different task. Our experiments include two popular pre-trained models: VGGFace and Resnet50. These models are able to classify gender and race easily with high performance (up to 95%) just adding a classification layer to the pre-trained model. Furthermore, suppressing most of the features from the features vector, they still classify gender and race quite well, showing the discriminative power of these attributes in VGGFace and Resnet50 pre-trained models. For our experiments, we use face images from the publicly available database LFW (Labeled Faces in the Wild). This database

comprises over 13000 face images with high variety of races, poses, ages and nationalities. We have chosen LFW database because many state-of-the-art face recognition algorithms use it as a benchmark for comparisons.

The rest of this paper is organized as follows: Section 2 explains the method proposed to achieve our goals. Section 3 describes the database and the experimental protocol. Section 4 presents the results and Section 5 summarizes the conclusions.

2 Proposed Method

Our objective is to measure to what extent gender and ethnicity information is present in the feature vectors generated by two of the most popular state-of-the-art face recognition models: VGGFace and Resnet50. To do this, we will follow two approaches:

Fixed classification: we employ the pre-trained models to extract the feature vector from each face image by removing the last classification layer from both models [5]. Then, we add a fully connected layer and train this layer in order to classify attributes. Finally, we will gradually suppress the most relevant features from the embedding layer (i.e., the layer that extracts the feature vectors) and test the networks for attribute classification and for identity verification as well. The idea is to test whether it is possible to keep a high performance in the verification task while suppressing the embedding features related to gender or ethnicity.

Retrained classification: the second approach studies what happens if we retrain the attribute classification layer after the suppression of these embedding features. To do this, in each iteration we suppress features and retrain the attribute classification layer using the remaining features.

3 Experiments

3.1 Database

The experiments are conducted in the LFW database (Labeled Faced in the Wild) [8]. LFW is one of the most popular datasets used in face recognition with more than 13000 face images of famous people collected from the web. We have used the aligned dataset where each image was aligned with funneling techniques [15] and labeled according to the gender, age and ethnicity among others (see [16] for details). The database was split into training and test set according to the "*View 1"* protocol [8]. This protocol employs up to 4038 face images for training and 1711 for testing. The database is highly biased, the statistics related to gender and ethnicity attributes are summarized in Table 1. We can observe that both gender and ethnicity distributions are unbalanced, as most of the images belong to Caucasian male. An unbalanced dataset could yield a drop of performance for classes with less samples, but as we will see in the next section the performance is stable across classes despite of the unbalanced dataset. We have decided to use the public protocol in order to allow a fair comparison with existing benchmarks.

3.2 Face recognition pre-trained models

We use two popular CNNs which have recently achieved some of the best state-of-the-art performance in face recognition tasks: VGGFace, proposed in [5], and ResNet50, proposed in [6].

VGGFace is a CNN with a VGG16 architecture (see [5] for details) trained from scratch with a dataset that contains more than 2,6 million images of 2622 celebrities (different from LFW). The architecture comprises 8 blocks of convolutional layers followed by activation layers like ReLU or maxpooling, and 3 blocks of fully connected layers with ReLU activations. VGGFace has an overall of 145,002,878 parameters split in 16 trainable layers (convolutional and fully connected layers). The results obtained testing VGGFace for verification with LFW and YFD (Youtube Face Dataset) datasets were 97.27% and 92.8% of accuracy respectively.

ResNet50 is another CNN based on Residual Neuronal Network architecture. This network is inspired in VGG nets, but with fewer filters and lower complexity. The key of this network is to insert shortcut connections among blocks which turn the network into a residual network version. ResNet50 has a total of 41,192,951 parameters split in 34 residual layers for training. In [17], they trained from scratch a ResNet50 network with VGGface2 dataset. VGGface2 contains up to 3,331 million images of 9,321 subjects and was collected with a huge range of pose, age and ethnicity. This Resnet50 model is aimed to improve the recognition performance over age and pose, achieving 98.0% of accuracy in verification testing with the IJB-A dataset [18] according to [17].

3.3 Experimental protocol

We employ the pre-trained models of VGGFace and ResNet50 provided in [19]. These model were tested using the "*unrestricted"* and *"outside training data*" protocols proposed in [8], due to both pre-trained models were trained with other databases, namely: [5] for VGGFace, and VGGFace2 [17] for Resnet50 respectively. The protocol includes 6000 one-to-one comparisons composed by 3000 genuine pairs (pairs of images from the same person) and 3000 impostor pairs (pairs of images belonging to different persons).

Identity verification: The feature vectors for each face image are extracted removing the last classification layer from both models. The number of features extracted is $L = 4096$ and $L = 2048$ for VGGFace and ResNet50 respectively. The distance between two face embeddings is obtained as the L2-distance for one-to-one comparisons in the verification task.

VGGFace achieves 92.3% of verification accuracy and Resnet50 84.1%. Note that performance on VGGFace is slightly worse as the one reported in [5]. This is due to different preprocessing of the data.

Gender and ethnicity classification: We add a fully connected layer with one unit in order to classify attributes (see Fig. 1) and freeze all remaining layers. Then, we train two separate layers for gender and ethnicity. About training details: the learning rate

Table 1. Distribution of gender and ethnicity in the considered LFW dataset.

	Gender (male)	Ethnicity (Cuacasian/Black/Asian)
Train	74.2%	79.8% / 6.3% / 13.9%
Test	74.4%	81.5% / 6.2% / 12.3%

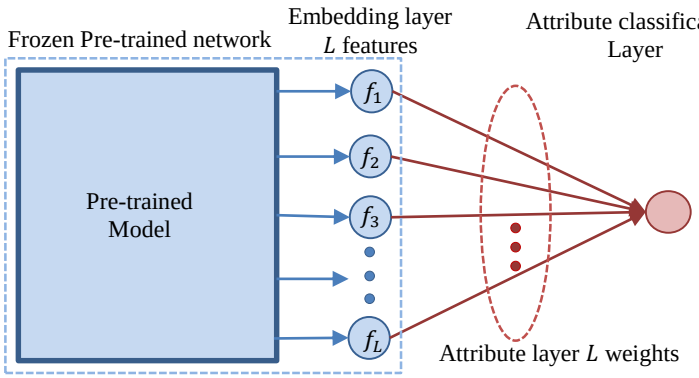

Fig 1. Architecture of the VGGFace network trained for gender classification. Only the $L = 4096$ weights from the attribute classification layer were trained from scratch, the remaining weights are equal to the pre-trained VGGFace model.

was $\alpha = 10^{-4}$, Adam optimizer was used with $\beta_1 = 0.9, \beta_2 = 0.999$ and $\varepsilon = 10^{-8}$ respectively, 50 epochs for VGGFace and 30 epochs for ResNet50 without minibatches. The activation function of the classification layer was sigmoid for gender (only one neuron for gender classification) and softmax for ethnicity (three neurons to classify among Black, Asian and Caucasian people). The performance will be reported in terms of correct classification accuracy.

4 Results

The performance achieved by both CNNs trained for classification of gender and ethnicity are summarized in Table 2. The best results are obtained with VGGFace network in both tasks.

Fig. 2 shows the distribution of weights obtained for the attribute-classification layer of VGGFace. We can observe similar normal distributions for both attributes with a high percentage of weights close to zero. There are two possible reasons for these weights: A) some features do not include information about these attributes and therefore to their contribution is minimized during the attribute classification training or; B)

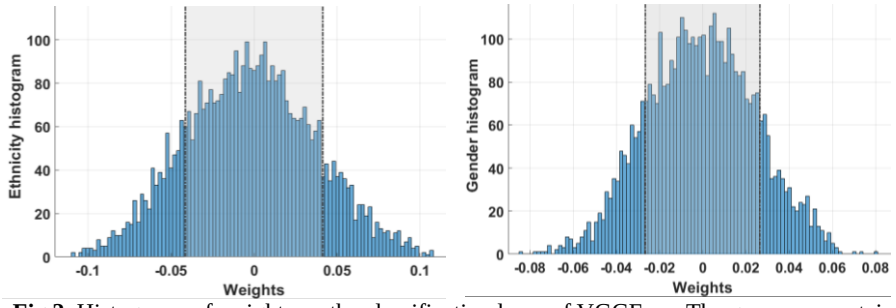

Fig 2. Histograms of weights on the classification layer of VGGFace. The grey area contains 70 percent of the weights.

Table 2. Accuracy (%) of VGGFace and ResNet50 trained for gender and ethnicity classification.

	VGGFace	ResNet50
Gender	94.8	89.01
Ethnicity	90.1	80.05

there is a high level of redundancy in the features and some features are minimized during training.

To measure the presence of gender or ethnicity information in the identity verification task, we will evaluate the verification performance decay by suppressing features from the embedding layer gradually in accordance of it gender/ethnicity importance. For that purpose, we suppress a feature by forcing the value of its weight in the classification layer to zero, starting from the weights with highest absolute values to lower ones. At the same time, we test the network for identity verification trying to keep a high performance in verification while we suppress the neurons related to gender or ethnicity discrimination.

Fig. 3 shows the performance decay for gender classification and identity verification tasks related to the percentage of suppressed features (from highest to lowest importance in gender classification). We can observe that both models are able to achieve almost 80% of accuracy for gender recognition with 50% of the features of the classification layer suppressed. Even though VGGFace achieves better gender classification performance, its performance drops faster than ResNet50, which is able to classify up to 75% of accuracy with 65% of weights suppressed. Regarding the verification task, it was surprising to find out that the performance keeps stable while we gradually suppress neurons from the embedding layer. In fact, VGGFace is able to verify with only 10% of the neurons (less than 400 neurons) from its embedding layer without performance decay. Regarding ResNet50, the performance starts to drop when 70% of neurons are suppressed. This behavior shows there is a very high redundancy of neurons in these networks in their embedding layer for identity verification. Finally, suppressing 80%, the two networks are still able to verify the identity of the users with a similar performance as using all their original embeddings.

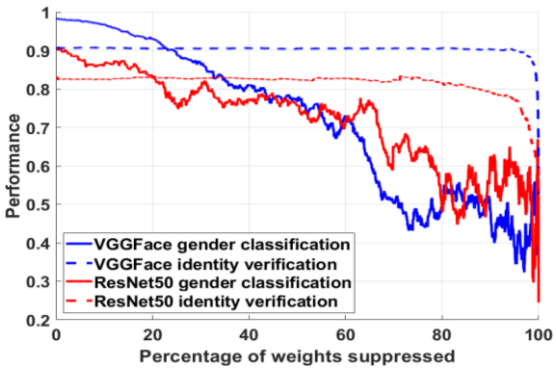

Fig 3. Classification and verification performance of both VGGFace and ResNet50 models according to the percentage of weights suppressed for gender classification.

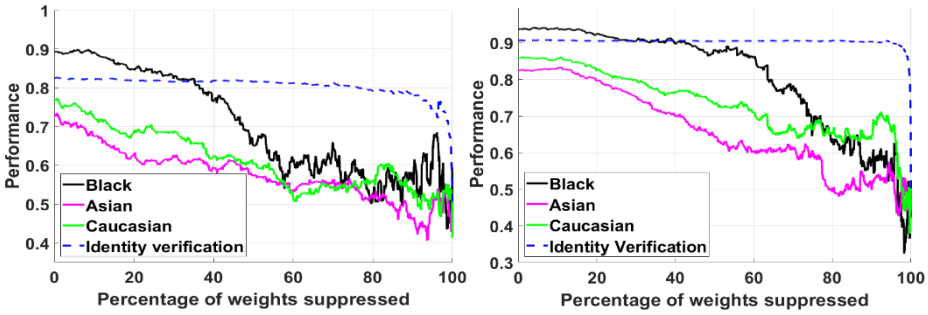

Fig 4. Classification and verification performance of ResNet50 (left) and VGGFace (right) according to percentage of weights suppressed for ethnicity classification.

In ethnicity classification, our experiments include three classes: Caucasian, Asian and Black. The classification accuracy was calculated for each class according to a one-vs-all protocol. In this case, the classification layer has $3 \times L$ weights (L weights for each class) and we add all three weights vectors in order to have only one vector of L weights to sort [20]. The idea is to sort the weights according to the most relevant weights for all classes (ethnicities).

Fig. 4 shows the performance decay for ethnicity classification in both models related to the number of weights suppressed (from highest to lowest importance in ethnicity classification). It is remarkable how well both networks classify Black with 90% of accuracy even if the training dataset was highly unbalanced. Regarding Caucasian and Asian, ResNet50 achieves worse performance and decays faster than VGGFace.

Regarding identity verification, we can see again that performance is not affected by the neurons suppressed until we suppress almost all of them. These experiments suggest that we can suppress features in order to reduce the gender and ethnicity bias while keeping the verification performance of the face recognition algorithms. However, the results showed in Figures 3 and 4 were obtained on the basis of a training phase including all the embedding features. What if we retrain the attribute classification layer after the suppression of these features? To do this, in each itineration we suppress again the

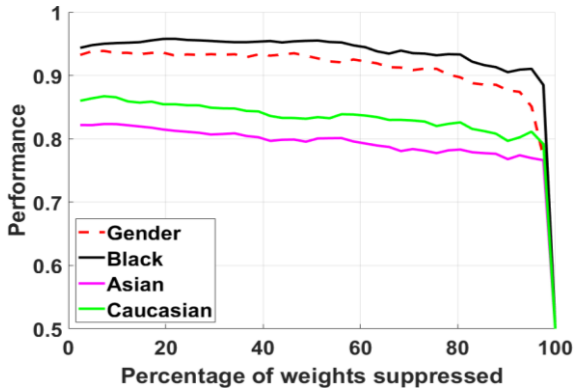

Fig 5. Performance of VGGFace for gender and ethnicity classification retraining the classification layer in each iteration.

most “relevant” feature (the one with the highest absolute weight value) of the embedding layer and we retrain the attribute classification model using the remaining features, always freezing the network up to the embedding layer. Fig. 5 depicts the performance for VGGFace. As we expected, the model is able to keep almost the same performance until 50% of features are suppressed. This demonstrates that gender and ethnicity attributes are latent in almost all features of the embedding layer for both pre-trained networks. Even with 85% of the features suppressed they are still able to classify quite well, revealing the discriminative power of these models.

5 Conclusions

In this paper, we have studied the importance of gender and ethnicity attributes in the decision-making of the most popular face recognition technologies. Although these attributes are useful for recognition, the risk using them for unethical purpose is tacit.

Firstly, we have explored how well VGGFace and Resnet50 pre-trained models classify gender and ethnicity by training a classification layer connected to the embedding layer of these models. The results suggest that these networks are able to discriminate among gender and ethnicity without performance decay suppressing almost 50% of the neurons of their embedding layer.

We have then studied the impact in identity verification of removing features related to gender and ethnicity. For the considered deep models, we have shown that removing up to 90% of the embedding features most related to gender and ethnicity does not affect much their performance in identity verification.

As future work, we are interested in more advanced methods for improving privacy in biometrics data processing, both at the learning stage [21] or by incorporating cryptographic constructions [22].

6 Acknowledgments

This work was funded by the project CogniMetrics (TEC2015-70627-R) and BioGuard (Ayudas Fundación BBVA a Equipos de Investigación Científica 2017).

References

1. Neves, J., Narducci, F., Barra, S., Proença, H.: Biometric recognition in surveillance scenarios: a survey. Artificial Intelligence Review, vol. 46, no. 4, pp. 515–541, (2016).
2. Gonzalez-Sosa, E., Vera-Rodriguez, R., Fierrez, J., Ortega-Garcia, J.: Exploring Facial Regions in Unconstrained Scenarios: Experience on ICB-RW, In: IEEE Intelligent Systems, vol. 33, n. 3, pp. 60–63, May (2018).
3. Selinger, E., Polonetsky, J., Tene, O.: The Cambridge Handbook of Consumer Privacy. 1st edn. Cambridge University Press (2018).
4. Ranjan, R., et al.: Deep Learning for Understanding Faces: Machines May Be Just as Good, or Better, than Humans. In: IEEE Signal Processing Magazine, vol. 35, pp. 66–83, (2018).

5. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Proceedings of the British Machine Vision, vol. 1, no. 3, p. 6 (2015).
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp.770–778, (2016).
7. Schroff, F., Kalenichenko, D., Philbin., J.: Facenet: A unified embedding for face recognition and clustering. In: Proc. CVPR (2015).
8. LFW Homepage, http://vis-www.cs.umass.edu/lfw/, last accessed 2018/6/15.
9. Tome, P., Fierrez, J., Vera-Rodriguez, R., Nixon, M.: Soft Biometrics and their Application in Person Recognition at a Distance. In: IEEE Transactions on Information Forensics and Security, vol. 9, no. 3, pp. 464–475, (2014).
10. Tome, P., Vera-Rodriguez, R., Fierrez, J., Ortega-Garcia, J.: Facial Soft Biometric Features for Forensic Face Recognition, In: Forensic Science International, vol. 257, pp. 171–284, December (2015).
11. Dantcheva, A., Elia, P., Ross, A.: What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics. In: IEEE Transactions on Information Forensics and Security, vol. 11, no. 3, pp. 441–467, March (2016).
12. Prabhakar, S., Pankanti, S., Jain, A.K.: Biometric recognition: Security and privacy concerns. In: IEEE Security Privacy Mag, vol.1, no. 2, pp. 33–42, (2003).
13. EU 2016/679 (General Data Protection Regulation), https://gdpr-info.eu/, last accessed 2018/10/17.
14. Goodman, B., Flaxman, F.: European Union regulations on algorithmic decision-making and a "right to explanation". In: AI Magazine, vol. 38, (2016).
15. Huang, G.B., Mattar, M., Lee, H., Learned-Miller, E.: Learning to Align from the scratch. In: Advances in Neural Information Processing Systems NIPS, (2012).
16. Gonzalez-Sosa, E., Fierrez, J. Vera-Rodriguez, R., Alonso-Fernandez, F.: Facial Soft Biometrics for Recognition in the Wild: Recent Works, Annotation and COTS Evaluation.In: IEEE Trans. on Information Forensics and Security, vol.13, no. 7, (2018).
17. Cao, Q., Shen, L., Xie, W., M. Parkhi, O., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: arXiv:1710.08092, (2017).
18. Klare, B., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Jain., A.K.: Pushing the frontiers of unconstrained face detection and recognition: Iarpa Janus benchmark A. In: CVPR, pp. 1931–1939, (2015).
19. Malli., R.C., Github Homepage, https://github.com/rcmalli/keras-vggface, last accessed 2018/6/15.
20. Fierrez, J., Morales, A., Vera-Rodriguez, R., Camacho, D.: Multiple Classifiers in Biometrics. Part 1: Fundamentals and Review. Information Fusion, vol. 44, pp. 57–64, (2018)
21. Mirjalili, V., Raschka, S., Namboodiri, A., Ross, A.: Semi-Adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images. In: Proc. of 11th IAPR International Conference on Biometrics, Australia, February (2018).
22. Gomez-Barrero, M., Maiorana, E., Galbally, J., Campisi, P., Fierrez, J.: Multi-Biometric Template Protection Based on Homomorphic Encryption. In: Pattern Recognition vol. 67, pp. 149–163, (2017).